STEREOLOGICAL ESTIMATION OF THE DISTRIBUTION OF SPHERICAL BODIES WITH CONSTANT RADIUS

by

FUNG Siu-kwong

A Thesis Submitted to the

Graduate School

of

The Chinese University of Hong Kong

(Division of Statistics)

In Partial Fulfillment

of the Requirements for the Degree of

Master of Philosophy (M. Phil.)

HONG KONG

June, 1995

THE CHINESE UNIVERSITY OF HONG KONG

GRADUATE SCHOOL

The undersigned certify that we have read a thesis, entitled "Stereological Estimation of the Distribution of Spherical Bodies with Constant Radius" submitted to the Graduate School by Fung Siu Kwong (馮 兆 光) in partial fulfillment of the requirement for the degree of Master of Philosophy in Statistics. We recommend that it be accepted.

Kim-Hung Li

Dr. K.H. Li,

Supervisor

Dr. K.H. Wu

Tai Sy Lau

Dr. T.S. Lau

Dr. S.K. Tse,

External Examiner

DECLARATION

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

ACKNOWLEDGEMENT

The author is deeply indebted to his supervisor, Dr. K.H. Li. for his generosity of encouragement and supervision. During the research program, Dr. Li gave me numerous comment and suggestions. Also, I would like to thank my colleagues for their kind assistance and my friends who gave me support and encouragement.

FUNG Siu-kwong

Department of Statistics

The Chinese University of Hong Kong

June, 1995.

Abstract

In stereology, the classical corpuscle problem has been studied at considerable length since the first solution by Wicksell(1925). Stereologists usually perform their inference under the assumption that the observed profiles are independently distributed. However, in practice, it is more likely that the objects will exert the attraction or repulsion on the others. Our interest puts emphasis on the dependence among the objects. For the ease of the computation and the comparison with the traditional method, we assume that the objects are constant radius spheres. We then use the algorithm proposed by Ripley (1977) to simulate the spatial distribution of the spheres. Moreover, we make use of the result from Gibbs Sampler to help us to find the Bayesian estimator of R, which is shown to be superior to the traditional method of moment estimator in terms of mean square error. Then, we further study the cases when (i) the number of the spheres in the specimen, N, is known, (ii) N is unknown or (iii) the boundary effect exists. Finally, we extent our consideration to spheres with varied radius.

Keywords : Stereology, Gibbs Sampler, Pairwise interaction process

Contents

List of Tables

Chapter 1

Introduction

Stereology is a body of mathematical methods relating three-dimensional parameters defining the structure to two-dimensional measurements obtainable on sections of the structure. It was initially developed in biology and material science as a quick way of analyzing three-dimensional solid materials, such as rock, living tissue, and metals, from information visible on a two-dimensional plane section through the material.

Nowadays, stereology has found many applications in areas including anatomy, cell biology, pathology, material science, mineralogy, metallurgy, ecology, geology and petrology, image processing, and computer graphics. It is not the aim of stereology to reconstruct an entire three-dimensional object. Typically, only a section or sample is taken, and their spatial position is not recorded. Stereology uses non-parametric techniques to estimate the geometrical parameters such as volume density, V_v.

The classical corpuscle problem in stereology has been studied at considerable

length since the first solution by Wicksell(1925). The problem is to express the distribution of the radius of the corpuscles in terms of the distribution of the radius of the circular contours found in the section plane. They assume that there are a large number of spherical corpuscles of different sizes in the opaque specimen. Then many different methods are proposed to estimate the size distribution of the substance from that of the profile. For example Keiding, Jensen and Ranek(1972) use the maximum likelihood method to estimate the size distribution of liver cell nuclei from the observed distribution in a plane section. Later, the other version of the classical corpuscle problem, the Swiss cheese problem occurs. The size distribution of randomly distributed spherical holes in an opaque medium is to be inferred from the size distribution of the circular holes observed in sections. Also, Keiding and Anderson(1992) show that the χ-distribution of Keiding, Jensen and Ranek(1972) may be used as the basis for a parametric maximum solution for the Swiss cheese problem.

Stereologists usually perform their inferences under the assumption that the observed profiles are independently distributed. However, in practice, it is more likely that the objects will exert the attraction or repulsion on the others. It is natural to take this effect into consideration. Our interest lays emphasis on the dependence among the objects. We suppose that the specimen is rectangular and the objects, which lie inside it, are spherical. Such a shape assumption is a severe restriction since objects are in practice often of complex and unknown shape, at least in biology, and may be even non-convex. Stereologists have developed some techniques to solve the arbitrarily shaped objects problem, such as Jensen (1987).

In our case, the reasons of the spherical assumption are as follows. First, many traditional analyses based on the assumption of sphere. In order to compare with those results, we adapt this. Second, it is easier for us to model the distribution of spheres.

The main purpose of this thesis is to consider the effect of the dependence among the objects. Firstly, we concentrate on the constant radius case. Then we will extend our method to the varied radius case, which will be the focus of future study. Throughout the discussion, the main interesting parameter that we want to estimate is the constant radius R.

In Chapter 2, we describe the traditional assumption, that is the independence assumption. We introduce an unbiased estimator of R, which is shown to be superior to the traditional method of moment estimator in terms of the mean square error MSE. Finally, we construct the confidence interval of R and the estimated variance for this unbiased estimator.

In Chapter 3, we concentrate on the effect of the dependence assumption on the estimator. Firstly, we consider the situation when N is known. We model the distribution of the spheres using Ripley's (1977) idea. We use Gibbs Sampler to estimate the posterior distribution of R without directly calculating it. Then, we illustrate our simulation procedure and demonstrate how to get the simulated parameters and the density. Finally, we relax our assumption and consider the situation when the number of spheres is unknown. Also, we consider the problem of the boundary effect.

In Chapter 4, we perform a simulation study to compare the results between

the independence and dependence case. We divide the simulation study into three parts: (i) N is known, (ii) N is unknown and (iii) the boundary problem. For each part, we also want to study the effect of the intensity level which is measured by the volumn density. Thus we further divide each part into three sections from the low intensity to the high intensity to investigate the intensity effect.

In Chapter 5, we make a conclusion and discuss a possible extension of our present setting. The study of the case of the dependence with constant radius is considered as a first step for the analysis of the case with varied radius. Hence, we simulate an artificial example to study the case of varied radius. Then we will compare the Bayesian estimator with the independent one, $\hat{R}_{MME}$, in terms of MSE.

Chapter 2

Independently distributed spheres with constant radius

Let us consider a rectangular specimen with some spherical bodies inside it. We assume that the bodies are independently distributed and have constant radius, R. This specimen is cut by a section plane and we get the profiles on the plane (profile means the circle which is seen on the section plane). Suppose we observe $n(n > 1)$ profiles with radius $x_1, x_2, ..., x_n$. Our objective is to estimate the constant radius R. The estimator used in stereology, which is for the general varied radius case, is in fact the method of moment estimator.

In Sections 2.1, and 2.2, we introduce the method of moment estimator and maximum likelihood estimator of R. In Section 2.3, we introduce an unbiased estimator of R, which is shown to be superior to the traditional method of moment estimator, and in the last section we calculate the confidence interval and the estimated variance for this unbiased estimator of R.

2.1 The method of moment estimator of R

Let $X_1, X_2, ..., X_n$, where X_i is the section radius of sphere i, be the random sample of size n from the specimen cut by a section plane. Due to the assumption of independence, we only consider the marginal distribution of X_i. Suppose we place the sphere into a $(x, y, z)-$ coordinate system with the sphere centre locating at the origin. Without loss of generality, we can assume the section plane to be parallel to the (x, y) plane. The section plane cuts the sphere randomly, i.e. $Z \sim U(-R, R)$, where Z is the distance of the random section plane from the origin. Then, the profile radius X forms a right-angled triangle with the sphere radius R and the distance Z, and satisfies the equation $X^2 + Z^2 = R^2$. We can obtain the density and distribution function of X by the transformation of the random variable Z. Denote them as $f_X(x)$ and $F_X(x)$ respectively.

$$f_X(x) = \frac{x}{R\sqrt{R^2 - x^2}}, \qquad R \geq x \geq 0. \tag{2.1}$$

$$F_X(x) = 1 - \frac{\sqrt{R^2 - x^2}}{R}, \qquad R \geq x \geq 0. \tag{2.2}$$

The expected value of X is

$$E(X) = \int_0^R \frac{x^2 dx}{R\sqrt{R^2 - x^2}} = \frac{R\pi}{4}.$$

The method of moment estimator of R is

$$\hat{R}_{\text{MME}} = \frac{4\bar{X}}{\pi},$$

where $\bar{X}$ is the sample mean of X_i's.

2.2 Maximum likelihood estimator of R

Under the same condition as that in the previous section, we have a random sample of size n with the section radius, $x_1, ..., x_n$. Then the likelihood function, $L(R \mid x_1, \cdots, x_n)$ is

$$\begin{aligned} L(R \mid x_1, \cdots, x_n) &= \prod_{i=1}^{n} f_X(x_i \mid R) \\ &= \frac{1}{R^n} \prod_{i=1}^{n} \frac{x_i}{\sqrt{R^2 - x_i^2}} I_{[0,R]}(x_i) \\ &= \frac{1}{R^n} \prod_{i=1}^{n} \frac{x_i}{\sqrt{R^2 - x_i^2}} I_{[x_{(n)},\infty)}(R). \end{aligned} \tag{2.3}$$

where $I_A(x)$ is the indicator function of the set A, and $X_{(n)}$ is the maximum of the n observed section radius. Thus the maximum likelihood estimator of R is

$$\hat{R}_{\text{MLE}} = X_{(n)}.$$

2.3 Mean squared errors of the estimators of R

At first, let us consider

$$\hat{R}_{MME} = \frac{4}{\pi}\bar{X}.$$

It is easy to show that $E(\hat{R}_{MME}) = R$. Thus $\hat{R}_{MME}$ is an unbiased estimator of R, and its variance is

$$Var(\hat{R}_{MME}) = \frac{16}{n\pi^2} Var(X).$$

Since $E(X^2) = \frac{2}{3}R^2$, and $E(X) = \frac{R\pi}{4}$, we have

$$Var(\hat{R}_{MME}) = (\frac{32}{3\pi^2} - 1)\frac{R^2}{n}.$$

Next we consider $\hat{R}_{MLE} = X_{(n)}$. Before evaluating the mean and variance of this estimator, we find the density function of $X_{(n)}$.

$$\begin{aligned} f_{X_{(n)}}(x) &= n f_X(x)[F_X(x)]^{n-1} \\ &= \frac{nx}{R\sqrt{R^2-x^2}}\left(1-\sqrt{\frac{R^2-x^2}{R}}\right)^{n-1} \qquad R \geq x \geq 0. \end{aligned} \tag{2.4}$$

$$\begin{aligned} E(X_{(n)}) &= \int_0^R \frac{nx^2}{R\sqrt{R^2-x^2}}\left(1-\sqrt{\frac{R^2-x^2}{R}}\right)^{n-1} dx \\ &= nR\int_0^{\frac{\pi}{2}} \sin^2\theta(1-\cos\theta)^{n-1}d\theta \\ &= 2^{n+2}nR\int_0^{\frac{\pi}{4}} (\sin^2\phi)^n - (\sin^2\phi)^{n+1}d\phi \\ &= HR, \end{aligned} \tag{2.5}$$

where $H = 2^{n+2}n\int_0^{\frac{\pi}{4}} (\sin^{2n}\phi - \sin^{2(n+1)}\phi)d\phi$.

Clearly,

$$MSE(X_{(n)}) = E(X_{(n)}^2) - R^2(2H-1).$$

In order to calculate the MSE of $X_{(n)}$, we need to calculate $E(X_{(n)}^2)$ and H. Consider

$$\begin{aligned} E(X_{(n)}^2) &= nR^2\int_0^{\frac{\pi}{2}} \sin^3\theta(1-\cos\theta)^{n-1}d\theta \\ &= nR^2\int_0^{\frac{\pi}{2}} (\cos^2\theta(1-\cos\theta)^{n-1} - (1-\cos\theta)^{n-1})dcos\theta \\ &= \frac{n(n+3)R^2}{(n+1)(n+2)}. \end{aligned} \tag{2.6}$$

Then we need to calculate the value H. Recall that

$$H = 2^{n+2}n\int_0^{\frac{\pi}{4}} [\sin^{2n}\phi - \sin^{2(n+1)}\phi]d\phi. \tag{2.7}$$

From the Mathematics Handbook (1979, p.265), we have

$$\int_0^{\frac{\pi}{4}} (\sin^{2n}\phi)d\phi = \frac{1}{\sqrt{2}} \sum_{r=0}^{n-1} \frac{(2n)!(r!)^2(\frac{1}{\sqrt{2}})^{2r+1}}{2^{2n-2r}(2r+1)!(n!)^2} + \frac{(2n)!\pi}{2^{2n}(n!)^2}. \tag{2.8}$$

Substituting (2.8) into (2.7) and after tedious simplification, we get

$$H = \frac{n}{n+1}\left(1 + \frac{(2n-1)!!}{n!}(\frac{\pi}{2} - \sum_{r=0}^{n-1} \frac{r!}{(2r+1)!!})\right), \tag{2.9}$$

where $(2n-1)!! \equiv 1 \cdot 3 \cdot 5 \cdots (2n-1)$. Due to the complexity of H, we want to get an approximate value.

Proposition 2.1 :

$$\frac{n(2n+5)}{(n+2)(2n+1)} \leq H \leq \frac{n(2n+3)}{(n+1)(2n+1)}. \tag{2.10}$$

Proof:

Firstly, we prove the following claim:

$$\frac{(2n+3)n!}{(n+2)(2n+1)!!} \leq \frac{\pi}{2} - \sum_{r=0}^{n-1} \frac{r!}{(2r+1)!!} \leq \frac{2(n!)}{(2n+1)!!}. \tag{2.11}$$

As n tends to infinity, the value of H tends to 1. So $\sum_{r=0}^{\infty} \frac{r!}{(2r+1)!!} = \frac{\pi}{2}$. Thus $\frac{\pi}{2} - \sum_{r=0}^{n-1} \frac{r!}{(2r+1)!!} = \sum_{r=n}^{\infty} \frac{r!}{(2r+1)!!}$.

It suffices to prove that

$$\frac{(2n+3)n!}{(n+2)(2n+1)!!} \leq \sum_{r=n}^{\infty} \frac{r!}{(2r+1)!!} \leq \frac{2(n!)}{(2n+1)!!},$$

which is equivalent to prove

$$\frac{(n+1)n!}{(n+2)(2n+1)!!} \leq \sum_{r=n+1}^{\infty} \frac{r!}{(2r+1)!!} \leq \frac{n!}{(2n+1)!!}.$$

Consider

$$\begin{aligned}\sum_{r=n+1}^{\infty}\frac{r!}{(2r+1)!!} &= \frac{(n+1)!}{(2n+3)!!}+\frac{(n+2)!}{(2n+5)!!}+\frac{(n+3)!}{(2n+7)!!}+\cdots \\ &= \frac{n!(n+1)}{(2n+3)(2n+1)!!}+\frac{n!(n+1)(n+2)}{(2n+5)(2n+3)(2n+1)!!}+\cdots \\ &= \frac{n!}{(2n+1)!!}\{\frac{(n+1)}{(2n+3)}+\frac{(n+1)(n+2)}{(2n+5)(2n+3)}+\cdots\}. \qquad (2.12)\end{aligned}$$

Define

$$\begin{aligned}K_n &\equiv \frac{(n+1)}{(2n+3)}+\frac{(n+1)(n+2)}{(2n+5)(2n+3)}+\cdots \\ &\equiv a_1+a_1a_2+a_1a_2a_3+\cdots\end{aligned}$$

where $a_1=\frac{n+1}{2n+3}, a_2=\frac{n+2}{2n+5}, a_3=\frac{n+3}{2n+7},\cdots$ As $a_1\le a_2\le a_3\cdots\le\frac{1}{2}$, it implies that $\frac{a_1}{1-a_1}\le K_n\le 1$. That is,

$$\frac{n+1}{n+2}\le K_n\le 1.$$

Hence,

$$\frac{(2n+3)n!}{(n+2)(2n+1)!!}\le\sum_{r=n}^{\infty}\frac{r!}{(2r+1)!!}\le\frac{2(n!)}{(2n+1)!!}.$$

It complete the proof of (2.11). Substitute (2.11) into (2.9), the result follows.

We use the following method to get a more accurate result. Let

$$H=\frac{n}{n+1}\left(1+\frac{(2n-1)!!}{n!}(\frac{\pi}{2}-\sum_{r=0}^{n-1}\frac{r!}{(2r+1)!!})\right)=\frac{n(2n+a+\epsilon_n)}{(n+1)(2n+1)} \qquad (2.13)$$

where a is a constant and ϵ_n tends to zero as n tends to infinity. Define

$$H_n=\sum_{r=0}^{n-1}\frac{r!}{(2r+1)!!}.$$

Rewrite (2.13) as

$$\sum_{r=0}^{n-1} \frac{r!}{(2r+1)!!} = \frac{\pi}{2} - \frac{n!}{(2n-1)!!}(\frac{2n+a+\epsilon_n}{2n+1} - 1). \tag{2.14}$$

Observe that

$$H_{n+1} = H_n + \frac{n!}{(2n+1)!!}. \tag{2.15}$$

Substituting (2.14) into (2.15), we get

$$\frac{\pi}{2} - \frac{(n+1)!}{(2n+1)!!}(\frac{2n+2+a+\epsilon_{(n+1)}}{2n+3} - 1) = \frac{\pi}{2} - \frac{n!}{(2n-1)!!}(\frac{2n+a+\epsilon_n}{2n+1} - 1) + \frac{n!}{(2n+1)!!}.$$

It follows that

$$\frac{(n+1)(a+\epsilon_{(n+1)}-1)}{(2n+3)} = a + \epsilon_n - 2.$$

As n tends to infinity, ϵ_n and $\epsilon_{(n+1)}$ tend to 0. We have $a = 3$.

Similarly, we can get more accurate result by replacing $a + \epsilon_n$ by $3 + \frac{b}{n} + \delta_n$ where $n\delta_n$ tends to zero as n tends to infinity. Repeat the above procedure, and solve the value b which is equal to -1.

Repeatedly, for more accurate results, we try $3 - \frac{1}{n} + \frac{c}{n^2} + \beta_n$ where $n^2\beta_n$ tends to 0 as n tends to infinity. Then we get $c = 3$. Certainly, we want our approximation as simple as possible. However, it should be accurate up to certain extent. Thus, in order to avoid using the square term of n, we consider

$$3 - \frac{1}{n+d} = 3 - \frac{1}{n} + \frac{3}{n^2} + o(\frac{1}{n^2})$$

where d is a constant which need to be determined. After simplification, we get $d = 3$. Hence, we decide to use the following approximation

$$H \approx \frac{n(2n+3-\frac{1}{n+3})}{(2n+1)(n+1)}.$$

Empirical checking shows that this approximation is correct up to the fourth decimal place.

Clearly $X_{(n)}$ is a biased estimator as it always underestimates the true value R. Before comparing their MSE, we find another estimator, which is a multiple of $X_{(n)}$. Consider $\hat{R} = \gamma X_{(n)}$, where γ is a constant. The corresponding MSE is,

$$\begin{aligned} MSE &= (E(\hat{R}) - R)^2 + Var(\hat{R}) \\ &= (\gamma H - 1)^2 R^2 + \gamma^2 Var(X_{(n)}). \end{aligned} \tag{2.16}$$

The choice of γ, say γ_n^* that minimizes the MSE is

$$\begin{aligned} \gamma_n^* &= \frac{H}{H^2 + \frac{Var(X_{(n)})}{R^2}} \\ &= \frac{(n+1)(n+2)H}{n(n+3)}, \end{aligned} \tag{2.17}$$

because $Var(X_{(n)}) = (\frac{n(n+3)}{(n+1)(n+2)} - H^2)R^2$.

Another choice of γ is to make it unbiased, that is $\gamma = \frac{1}{H}$. Now, we have three estimators. They are $X_{(n)}, \gamma_n^* X_{(n)}$ and the unbiased estimator $X_{(n)}/H$. We notice that for the value of n larger than 9, their difference are within one percentage. For ease of the comparison with the traditional estimator, we simply consider the unbiased estimator, $X_{(n)}/H$. We denote it as $\hat{R}_{adjMLE}$. Then, we compare $\hat{R}_{adjMLE}$ with the traditional estimator of R, i.e. $\hat{R}_{MME}$. As they are both unbiased, we just need to compare their variances. The following proposition shows that the variance of $\hat{R}_{adjMLE}$ is smaller than that of $\hat{R}_{MME}$. Besides, it shows that $Var(\hat{R}_{MME}) = O(\frac{1}{n})$ and $Var(\hat{R}_{adjMLE}) = O(\frac{1}{n^3})$. When n is small, we write a program to evaluate the value of H and get the estimates and variances of $\hat{R}_{MME}$ and $\hat{R}_{adjMLE}$. We list all these values in Table 2.1.

Proposition 2.2 : For $n \geq 1, Var(\hat{R}_{adjMLE}) \leq Var(\hat{R}_{MME})$. The equality holds only when $n = 1$.

Proof: Recall

$$\begin{aligned} Var(\hat{R}_{MME}) &= (\frac{32}{3\pi^2} - 1)\frac{R^2}{n}. \\ Var(\hat{R}_{adjMLE}) &= (\frac{n(n+3)}{(n+1)(n+2)H^2} - 1)R^2 \\ &\leq (\frac{(n+2)(2n+1)^2(n+3)}{n(n+1)(2n+5)^2} - 1)R^2 \\ &= \frac{2(n+3)R^2}{n(n+1)(2n+5)^2} \end{aligned}$$

Then to prove the proposition is equivalent to prove

$$\frac{32}{3\pi^2} - 1 - \frac{2(2n+3)}{(n+1)(2n+5)^2} \geq 0.$$

After simplification, the left hand side is equal to

$$\frac{4an^3 + 24an^2 + (45a-4)n + 25a - 6}{(n+1)(2n+5)^2},$$

where $a = \frac{32}{3\pi^2} - 1$. Let

$$f(n) = 4an^3 + 24an^2 + (45a-4)n + 25a - 6,$$

and for $n \geq 2$,

$$\frac{df(n)}{dn} = 12a(n+2)^2 - 3a - 4 \geq 0.$$

Also, $f(2) = 5.624$. Hence, the result follows.

For $n = 1$, $E(X_{(n)}) = \frac{R\pi}{4}$. It implies that $H = \frac{\pi}{4}$. Then it is easy to show that the two variances are equal.

2.4 Confidence interval of R

Suppose we have a data set $x_1, ..., x_n$, where x_i corresponds to the section radius of sphere i. It is found that $\hat{R}_{adjMLE}$ is better than $\hat{R}_{MME}$, in terms of MSE. In practice, after we have obtained the data, we only have to find out the maximum section radius and multiply it by the factor $\frac{1}{H}$ which will then become the desired estimate. For $n = 1, ..., 20$, we have tabulated the value of this factor, $\frac{1}{H}$ in Table 2.1. For $n(\geq 20)$ is large, we adapt the following approximation

$$H \approx \frac{n(2n + 3 - \frac{1}{n+3})}{(2n+1)(n+1)}.$$

The estimated variance of $\hat{R}_{adjMLE}$ is

$$\hat{Var}(\hat{R}_{adjMLE}) = \left(\frac{n(n+3)}{(n+1)(n+2)H^2} - 1 \right) \hat{R}^2_{adjMLE}.$$

We can construct a confidence interval of R. Consider

$$P\{X_{(n)} \leq R \leq \beta X_{(n)}\} = 1 - \alpha,$$

where $\beta(> 1)$ is a constant. Recall that the distribution function of $X_{(n)}$ is

$$F_{X_{(n)}}(x) = \left(1 - \frac{\sqrt{R^2 - x^2}}{R} \right)^n, \quad R \geq x \geq 0.$$

Then

$$\begin{aligned} P\{R \geq X_{(n)} \geq \frac{R}{\beta}\} &= 1 - \alpha. \\ -F_{X_{(n)}}(\frac{R}{\beta}) + F_{X_{(n)}}(R) &= 1 - \alpha. \\ 1 - F_{X_{(n)}}(\frac{R}{\beta}) &= 1 - \alpha. \\ \left(1 - \frac{\sqrt{\beta^2 - 1}}{\beta} \right)^n &= \alpha. \end{aligned} \tag{2.18}$$

Therefore, the $100(1-\alpha)$ percentage confidence interval for the R is $(X_{(n)}, \beta X_{(n)})$ where

$$\beta = \sqrt{\frac{1}{\alpha^{\frac{1}{n}}(2-\alpha^{\frac{1}{n}})}}.$$

Also, we list the β value for different levels of α in the rightmost columns of Table 2.1.

n	$\frac{1}{H}$	app.$\frac{1}{H}$	$\frac{Var(\hat{R}_{adjMLE})}{R^2}$	$\frac{Var(\hat{R}_{MME})}{R^2}$	β		
					$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.025$
1	1.27324	1.26316	0.81d-01	0.81d-01	2.2942	3.2026	4.5004
2	1.10604	1.10294	0.19d-01	0.40d-01	1.3704	1.5867	1.8530
3	1.05793	1.05660	0.73d-02	0.270d-01	1.1844	1.2898	1.4152
4	1.03686	1.03618	0.34d-02	0.20d-01	1.1122	1.1768	1.2528
5	1.02562	1.02524	0.18d-02	0.16d-01	1.0759	1.1202	1.1723
6	1.01889	1.01866	0.11d-02	0.13d-01	1.0550	1.0875	1.1257
7	1.01452	1.01437	0.66d-03	0.12d-01	1.0418	1.0667	1.0962
8	1.01152	1.01142	0.43d-03	0.10d-01	1.0328	1.0527	1.0761
9	1.00937	1.00930	0.30d-03	0.90d-02	1.0265	1.0427	1.0618
10	1.00777	1.00772	0.21d-03	0.81d-02	1.0218	1.0353	1.0513
11	1.00655	1.00651	0.15d-03	0.73d-02	1.0183	1.0297	1.0432
12	1.00560	1.00557	0.12d-03	0.67d-02	1.0156	1.0253	1.0370
13	1.00484	1.00482	0.88d-04	0.62d-02	1.0134	1.0219	1.0320
14	1.00423	1.00421	0.68d-04	0.58d-02	1.0117	1.0191	1.0280
15	1.00372	1.00371	0.54d-04	0.54d-02	1.0103	1.0168	1.0246
16	1.00331	1.00329	0.43d-04	0.50d-02	1.0091	1.0149	1.0219
17	1.00295	1.00295	0.35d-04	0.48d-02	1.0081	1.0133	1.0196
18	1.00266	1.00265	0.29d-04	0.45d-02	1.0073	1.0120	1.0176
19	1.00240	1.00240	0.24d-04	0.43d-02	1.0066	1.0108	1.0159
20	1.00218	1.00218	0.21d-04	0.40d-02	1.0060	1.0098	1.0145

Table 2.1: Values of $\frac{1}{H}$, Varinace of the two estimators; β value with $\alpha = 0.1, 0.05, 0.025$

Chapter 3

Dependently distributed spheres with the constant radius

In Chapter 2, we have discussed the method of estimating R under independence assumption. In this chapter, we change this assumption to the dependence one.

In the independent case, we assume that the spherical bodies are independently distributed among the specimen. This specimen is cut by a section plane to produce the profiles. It is equivalent to hold a sphere and cut it randomly. But now, due to the existence of the interaction among the spheres, we need to model the spatial distribution of spheres. In general, it is difficult to model a three-dimensional structure explicitly. Instead, stereologist uses the non-parametric techniques to estimate geometrical parameters such as volumn density. However, if we can assume a constant radius R and the spherical shape of the opaque objects, then we can model the distribution of spheres by the idea of Ripley(1977). Hence, we can utilize the results from the Gibbs Sampler to help us to estimate

R and the other parameters, such as V_v.

Throughout the following discussion, we define some notations. Seen spheres mean the n spheres which are cut by the section plane, while unseen spheres are not. We use U, which is an $n \times 1$ random vector $(u_1, \cdots, u_n)'$, to denote the location of the centre of the seen spheres relative to the section plane. For $i = 1, \cdots, n$, if $u_i = 1$, it means that the centre of the i th sphere lies above the section plane; if $u_i = -1$, it indicates that the centre of the i th sphere lies below the section plane. We use C, to denote the centres of unseen spheres, which is an $(N - n) \times 3$ random matrix $(c_{(n+1)}, \cdots, c_N)'$, where $c_j = (x_j, y_j, z_j)'$ for $j = n + 1, \cdots, N$, R to denote the radius of the sphere, and X_{obs} to denote the observed profiles. Finally, we use S_{com} to represent (X_{obs}, U, C).

In the coming sections, Section 3.1 introduces the results from the Gibbs Sampler which are relevant to our problem; Section 3.2 introduces the spatial distribution of the sphere; Sections 3.3 and 3.4 discuss the simulation algorithm and the explanation of the algorithm; Section 3.5 mentions the estimation of the posterior distribution of R; and the last two Sections, Sections 3.6 and 3.7 discuss the problem of unknown number of spheres in the specimen and the boundary problem.

3.1 Gibbs Sampler

Gibbs sampler is a technique for generating random variables from a (marginal) distribution indirectly, without having to calculate the density. Let us first explore

it in three variables, C, R, U case. Starting from these three variables, Gibbs sampler generates a sample of C, R, U from their conditional densities $f(c \mid r,u), f(r \mid c,u), f(u \mid c,r)$ instead of from their joint density. This is done by generating a 'Gibbs Sequence' of random variables, say

$$R_0 = r_0, U_0 = \vec{u}_0, C_1 = c_1, R_1 = r_1, U_1 = \vec{u}_1, \cdots, C_P = c_P, R_P = r_P, U_P = \vec{u}_P.$$

The initial values $R_0 = r_0$ and $U_0 = \vec{u}_0$ are specified. The rest is obtained iteratively by alternatively generating value from

$$\begin{aligned} C_{j+1} &\sim f(c \mid R = r_j, U = \vec{u}_j) \\ R_{j+1} &\sim f(r \mid C = c_{j+1}, U = \vec{u}_j) \\ U_{j+1} &\sim f(u \mid C = c_{j+1}, R = r_{j+1}) \end{aligned} \tag{3.1}$$

for $j = 1, 2, \cdots$. The distribution of R_P converges in distribution to $f(r)$, the true marginal distribution of R, as P tends to infinity. Thus when P is large enough, the final observation, namely $R_P = r_P$ is effectively a sample point from the marginal density of R.

Gibbs Sampling can also be used to estimate $f(r)$ by averaging the final conditional densities from each Gibbs sequence. Suppose we generate n 'Gibbs Sequences' of random variables, say

$$\begin{aligned} &R_{10} = r_{10}, U_{10} = \vec{u}_{10}, \quad \cdots, \quad C_{1P} = c_{1P}, R_{1P} = r_{1P}, U_{1P} = \vec{u}_{1P} \\ &R_{20} = r_{20}, U_{20} = \vec{u}_{20}, \quad \cdots, \quad C_{2P} = c_{2P}, R_{2P} = r_{2P}, U_{2P} = \vec{u}_{2P} \\ &\qquad\qquad \vdots \\ &R_{n0} = r_{n0}, U_{n0} = \vec{u}_{n0}, \quad \cdots, \quad C_{nP} = c_{nP}, R_{nP} = R_{nP}, U_{nP} = \vec{u}_{nP} \end{aligned} \tag{3.2}$$

Then, the marginal density function of R is estimated by

$$\hat{f(r)} = \frac{1}{n}\sum_{i=1}^{n} f(r \mid \vec{u}_{iP}, c_{iP}). \tag{3.3}$$

The theory behind is that the expected value of the conditional density is

$$E(f(r \mid U, C)) = \int\int f(r \mid u, c) f(u, c) du dc = f(r).$$

In the consequent discussion, we will concentrate on the iterative procedure mentioned in (3.1) and the estimation of the density in (3.3).

3.2 Model based on simulating spatial patterns

Firstly we list all the assumptions as follows.

[1] The specimen is a rectangular shape with length, l, width, w and height h.

[2] There are N spherical opaque objects whose centres lie inside the specimen.

[3] The specimen does not contain any part of sphere with centre lying outside the specimen.

[4] The section plane is parallel to the (x, y) plane with height fixed at $h/2$ measured from the bottom of the specimen.

In the modelling of the spatial structure, we apply the idea of Ripley(1977).

Model

The number of spheres in the specimen N is assumed to be known. Later, we will relax this assumption. Let the (x,y,z)-coordinate of the centre of the sphere

i be $S_i, i = 1, ..., N$. The joint density $f(s_1, ..., s_N \mid R)$ is taken to be a member of pairwise interaction process, i.e.

$$f(s_1, ..., s_N \mid R) \propto \prod_{i \neq j} k\{d(s_i, s_j)\} \tag{3.4}$$

where $d(s_i, s_j)$ is the distance between the centres of the i th and j th sphere. For the choice of $k(d)$, we can roughly divide it into three classes. The first class is

$$k_1(d) = \begin{cases} 0 & \text{for } 0 \leq d < 2R \\ 1 & \text{for } d \geq 2R \end{cases} \tag{3.5}$$

This class imposes a minimum distance $2R$ between the events. The second class is

$$k_2(d) = \begin{cases} 0 & \text{for } 0 \leq d < 2R \\ \frac{d-2R}{d_1-2R} & \text{for } 2R \leq d < d_1 \\ 1 & \text{for } d \geq d_1 \end{cases} \tag{3.6}$$

This class additionally discourages pairs of events less than d_1 apart and thereby produces a more regular pattern events than the first one. The third class is

$$k_3(d) = \begin{cases} 0 & \text{for } 0 \leq d < 2R \\ d_2 - \frac{d_2(d-2R)}{d_1-2R} & \text{for } 2R \leq d < d_1 \\ 1 & \text{for } d \geq d_1. \end{cases} \tag{3.7}$$

It involves an element of attraction between events a distance 2R to d_1 apart, giving a visual impression of aggregation.

For the simplicity of the simulation, we only consider $k_1(d)$ as our choice of $k(d)$ throughout the paper. Clearly

$$f(s_1 \mid s_2, ..., s_N, R) \propto \prod_{1<j} k_1\{d(s_j, s_1)\}.$$

Algorithm

For the initial step, the spheres are arbitrarily placed in the specimen. Then repeating the following steps: one point is chosen at random and then replaced by a point drawn from the conditional density of the remaining $N-1$ points. However, in our case, we have no idea of where the spheres are. We only have the observed profiles, i.e. X_{obs}. Therefore, in order to have a starting pattern, we need to simulate the positions of the spheres given X_{obs}. That means condition on the X_{obs}, we need to simulate the relative position of the centres of the seen spheres, U and the centres of the unseen spheres, C. Then we follow the above algorithm by replacing each sphere in turn rather than chosen at random. It does not affect the result of simulating the spatial distribution. Moreover, the theory shows that the process converges to the true distribution no matter what the starting distribution is. This algorithm is the same as that resulted from the Gibbs sampler.

Relation between R and (U, C)

In our problem, the unknown parameters are R, U, and C. Suppose we do not have any information about R, and it is intuitive that the smaller the value of R, the smaller chance the spheres overlap. We believe that conditional on the position of the centres of the spheres, R should be uniformly distributed, provided that the spheres do not overlap. For the distribution of the centres, U and C, we believe that conditional on R, the centres are uniformly distributed as far as they do not overlap. It is because after imposing the dependence among the specimen,

the spheres are attracted or repulsed by the others. Thus it is not reasonable that the spheres are uniformly distributed among the specimen. In short, we believe that the distribution of $R \mid (U, C)$ is uniform on the condition that the spheres do not overlap. The distribution of $(U, C) \mid R$ is uniform on the condition that the spheres do not overlap. It can be shown that the choices of the above conditional densities are consistent.

3.3 Algorithm

We give the main steps of the algorithm. More detail will be explained in the next section.

[1] Conditional on the observed profiles, say X_{obs}, we find the lower and upper bound of R, say R_l and R_u respectively. We use the maximum profiles radius, $R(1)$ as our initial value of R.

[2] Initialize the variable U.

[3] Deterministically locate the unseen sphere C.

[4] Repeat [5] - [7] P times to get one sequence of P values of R.

[5] Given R and U, simulate C.

[6] Given C and U, simulate R.

[7] Given R and C, simulate U.

[8] Calculate the posterior distribution of $R \mid S_{com}$, say $f_j(R \mid S_{com})$.

[9] Repeat [1]-[8] Q times to obtain Q different posterior distributions of $R \mid S_{com}$ and take average of them to estimate the true posterior distribution of $R \mid X_{obs}$, i.e. $f(R \mid \hat{X}_{obs})$ = average of $\{f_j(R \mid S_{com})\}_{j=1}^{Q}$.

Let us elaborate each step in more detail.

3.4 Explanation of the algorithm

STEP 1 (Find the lower and upper bounds of R)

Suppose we are given a data set $(a_i, b_i), x_i, i = 1, \cdots, n$, where n is the number of the observed profiles, x_i is the profile radius of sphere i, and (a_i, b_i) is the (x, y) coordinate of the centre of profile i. Since all the centres of the spheres lie within the specimen, feasible value of R must be bounded. A simple choice of R_l, say $R(1)$, is

$$R(1) = \max_{1 \le i \le n}(x_i).$$

We use it as the starting value of R. For the choice of R_u, we consider

$$R_u = \begin{cases} \sqrt{max(Z, h-Z)^2 + R(1)^2} & \text{if } N = n \\ max(Z, h-Z) & \text{if } N > n \end{cases} \tag{3.8}$$

where Z is the distance of the section plane from the bottom of the specimen.

STEP 2 (Locate the centres of seen spheres)

Assume that $S_1, ..., S_N$ have a density function

$$f(s_1, ..., s_N \mid R) \propto \prod_{i \neq j} k_1\{d(s_i, s_j)\},$$

where the first n spheres respresent the seen spheres and the rest are unseen. Given R, the centres of seen spheres are completely determined by U. From the density of f, the spheres are uniformly distributed among the specimen. As a consequence, the centres of seen spheres are equally likely to lie above or below the section plane as far as the spheres do not overlap each other. We try all possible arrangements for these seen spheres to place them into the specimen.

STEP 3 (Set the unseen spheres deterministically.)

For simplicity, we set the initial locations of the unseen spheres deterministically. We arrange the unseen spheres regularly and far away from the seen spheres to avoid overlapping.

STEP 4 (Given R and U, simulate C.)

Consider

$$f(s_j \mid s_{-j}, R) \propto \prod_{i \neq j} k_1\{d(s_i, s_j)\} \tag{3.9}$$

$j = n+1, \cdots, N$, where the notation of s_{-j} means that $\{s_1, \cdots, s_N\} \backslash \{s_j\}$. We pick an unseen sphere, s_j, one by one from $j = n+1$ to $j = N$, and replace it by the new s_j which is simulated from (3.9). In order to enhance the efficiency, we may improve our algorithm by using the following proposition.

Proposition 3.1:

Set S_{j0} be the current location of sphere j.

Suppose $(S_{j0} \mid S_1, \cdots, S_{j-1}, S_{j+1}, \cdots, S_N)$ follows the uniform distribution, $U(D)$, where D is the feasible region of S_j given S_{-j}. For $i \geq 1$, V_i, is a 3×1

random vector $(v_{i1}, v_{i2}, v_{i3})'$ with distribution symmetric at zero, and is independent of $S_{j,(i-1)}$. Set

$$S_{ji} = (S_{j,(i-1),1} + v_{i1} \pmod{l}, S_{j,(i-1),2} + v_{i2} \pmod{w}, S_{j,(i-1),3} + v_{i3} \pmod{h}). \tag{3.10}$$

We simulate S_{j1}, S_{j2} and so on, until it falls in D. Then the resulting S_{ji} is uniform on D.

From the above proposition, S_{j0} and S_{ji} represent the original and the new centre respectively. The idea of the proposition is that if the new simulated centre lies outside the specimen, we replace it back to the specimen by the transformation (3.10). If the newly simulated sphere overlaps with the others, we repeat the above process until we get the one which does not overlap with the others. After we have imposed this condition, the conditional distribution of the new centre is still uniform. For the proof of the proposition, we just point out the idea. Suppose the original centre is S_{j0}. Then we simulate the new one, say S_{j1}, and so on, until K times. As $f(S_{ji} \mid S_{j(i-1)}) = f(S_{j(i-1)} \mid S_{ji})$, $i = 2, \cdots, K$, the probability of the occurrence of the path $S_{j0} \to S_{j1} \to \cdots \to S_{jK}$ is the same as that of $S_{jK} \to S_{j(K-1)} \to \cdots \to S_{j1} \to S_{j0}$. Thus, the invariant property of the above process is preserved.

STEP 5 (Given C and U, simulate R.)

Given the position of the spheres, and X_{obs}, i.e. S_{com}, we simulate R from its conditional distribution which is uniform. Let R_o be the old R value. The

procedure is as follows.

1. Set $\alpha = R_l$ and $\beta = R_u$.

2. Simulate $T \sim U(\alpha, \beta)$. Check whether we can accept this or not (acceptance means substituting this T as a true value of R does not cause the overlapping of the spheres).

3. If the simulated T is accepted, the candidate of R is T, and exit.

4. If the simulated value is rejected, then basing on this information, we restrict our next value of T to a narrower interval instead of the original one. In the process, the rule for updating the interval of T is as follows. Suppose the simulated T is rejected. If T is greater than R_o, then we replace β by T. Otherwise we replace α by T. Go to (2).

STEP 6 (Given R and C, simulate U.)

The conditional distribution of U is uniform, as

$$f(s_j \mid s_{-j}, R) \propto \prod_{i \neq j} k_1\{d(s_i, s_j)\}, j = 1, \cdots, n.$$

The (x_j, y_j) coordinate of the j th seen sphere is known to be (a_j, b_j). Thus, given R, only the z coordinate of S_j is unknown and it is completely determined by u_j.

The procedure of simulating U is as follows. For $j = 1, \cdots, n$, it is equally likely for the centre of the seen sphere to lie above or below the section plane provided that the spheres do not overlap. Consequently, there is 0.5 chance to change the value of u_j.

[1] Generate a $U(0,1)$ random number, say v.

[2] If $v \leq 0.5$, keep the old u_j value and exit.

[3] if $v > 0.5$, keep the old u_j value if and only if changing the u_j value will lead to the overlapping of the new sphere with the existing spheres.

We iterate $P(= 10)$ times to get the simulated C, R and U values and we denote $R(P)$ to be the last simulated value of R. For the choice of P, we perform a simulation study in Chapter 4 to investigate the effect of P to the algorithm.

3.5 Posterior distribution of R

The posterior distribution of R is uniformly distributed in an interval. Hence, we need to determine the lower and upper limits of the interval. Two methods will be introduced to overcome this problem.

I. Analytical Method

Due to the non-overlapping of the spheres, the distance between any two centres in the specimen must be greater than or equal to 2R. That is

$$distance(s_i, s_j) \geq 2R \quad for \ \ all \ i \neq j.$$

We only need to solve the inequalities by substituting the values of s_i. The above inequalities may be difficult to be solved because for the seen sphere, s_i, $i = 1, \cdots, n$ is a function of R. Thus we propose the second method as below.

II. Bisection Method

We first illustrate how to find the upper limit of R. Since $R_l \leq R \leq R_u$, and $R(P)$ is the last simulated value of R, the upper limit of R should be within $(R(P), R_u)$.

The procedure to find the upper limit includes four steps.

[1] Set $\alpha = R(P)$ and $\beta = R_u$.

[2] Set $R = R_u$ and check whether the spheres overlap or not. If they are not overlapping (i.e. we accept this value of R), the upper limit of the posterior distribution of R is R_u; otherwise go to [3].

[3] Let $R_{bis} = (\alpha + \beta)/2$. Then we check whether we can accept or reject this value, $R = R_{bis}$. If we accept it, replace α by R_{bis}; otherwise, replace β by R_{bis}.

[4] If the difference between α and β is smaller than a preassigned value, stop and set the upper limit to be $(\alpha + \beta)/2$; otherwise go to [3].

Similarly, we can apply this method to find the lower limit of R. In that case, we need to set $\beta = R(P)$ and $\alpha = R_l$. Then go through the similar steps. Suppose we have got the desired results. In order to estimate the true posterior distribution of $R \mid X_{obs}$, we iterate the above algorithm $Q(= 20)$ times and get Q number of the posterior distributions, $\Pi_j(R \mid S_{com})$, $j = 1, \cdots, Q$. Consequently, the true one is estimated by

$$\hat{\Pi}(R \mid X_{obs}) = \frac{\sum_{j=1}^{Q} \Pi_j(R \mid S_{com})}{Q}$$

Then we can use the estimated posterior mean, say $\hat{R}_{bay}$, to estimate the constant radius R. The posterior mean can be obtained using either of the following two methods.

[I] We calculate the estimated posterior distribution, and then find its mean.

[II] For each simulated data set, the conditional posterior distribution of R is $R \sim U(a, b)$, say. Then $E(R) = (a+b)/2$ and $Var(R) = (b-a)^2/12$. Thus, the desired posterior mean and variance can be found by making use of the conditional expectation equation.

$$\begin{aligned} E(R \mid X_{obs}) &= E(E(R|X_{obs}, U, C)) \\ Var(R \mid X_{obs}) &= Var(E(R|X_{obs}, U, C)) + E(Var(R|X_{obs}, U, C)) \end{aligned}$$

Finally, for other interesting parameters, such as V_v, we can estimate it directly as we have the approximated posterior of R.

3.6 The number of spheres is unknown

In practice, it is more likely that N is unknown. Then our algorithm discussed before will not work. To overcome this, we have to modify our algorithm to cope with this problem. We have the following suggestion.

Suppose the specimen is cut randomly by a section plane. We can count the number of the profiles, say n, and find the maximum profile radius. For the centres of the seen spheres, they must lie within R distance above or below the section plane. Imagine that the section plane enlarges equally on each side to

become a slice of section with thickness $2R$. It will cover all the centres of the seen spheres. Based on it, N can be estimated in proportion through the following equation.

$$\hat{N} = \frac{n}{2\hat{R}A_s}V \tag{3.11}$$

where A_s is the section area; V is the volume of the specimen. To use this estimator of N, we should decide the estimator of R. If we just use the current simulated value of R, then N will change in every iteration. Therefore, we decide to use $\hat{R} = \hat{R}_{adjMLE}$. Indeed, this choice of estimator of N is reasonable and simple. We just use the idea of proportion to find this estimator. We will discuss other choices of the estimator of N in the last chapter. At present, we will use the estimator in (3.11) throughout all the simulation when N is unknown.

3.7 Boundary Effect

In our discussion, we ignore the effect of the other spheres whose centres are outside the specimen. Since the outside spheres also attract or repulse the inside one, we should take them into consideration. Obviously, the effect will diminish as the outside spheres lie far away from the section plane. We need to know how far will this effect be insignificant. Therefore, we will conduct a simulation study to investigate this effect. The behaviour of this factor will be discussed in Chapter 4.

Chapter 4

Simulation Study

In Chapter 2, we have discussed the case of independently distributed spheres and derived an estimate of R, say $\hat{R}_{adjMLE}$, which is shown to be superior to the method of moment estimator. In Chapter 3, we relax this independence assumption to the condition of dependence. Also, we have derived the estimate of R, say $\hat{R}_{bay}$ under different conditions, such as unknown N. In this chapter, Section 4.1 determines the convergence of the Gibbs Sequence which is mentioned in (3.1). In the rest of the sections, we compare the results of the adjusted maximum likelihood estimator, $\hat{R}_{adjMLE}$ and the Bayesian estimator under different conditions. They are (i) N is known, (ii) N is unknown and (iii) the boundary problem. Also, in order to study the effect of varied intensity, we further divide the simulation into three parts, from low to high intensity.

4.1 Determination of the convergence of the Gibbs sequence

Recall that in Section 3.1, if P is large enough, the final observation of $R_P = r_P$ is effectively a sample point from the marginal density of R, $f(r)$. We need to determine P such that the estimated posterior mean of R will close to the true one. The procedure is as follows.

We simulate $N = 10$ spheres with constant radius $R = 0.12$ in a unit cubic specimen by using the algorithm stated in Section 3.2. We fix the section plane in the middle of the specimen. That is, if we place the specimen into the origin of the $(x, y, z)-$ coordinate system, the height of the section plane is $1/2$. Then, we get the X_{obs}. We assume that N is known to be 10 throughout this simulation. In order to investigate the effect of the value of P only, we keep the number of times of simulation, Q, to 20. Then we vary P from 10 to 100. Since the simulated radius, R, is affected by the data, in order to eliminate this nuisance factor, we perform a randomized complete block design. The setting is as follows.

Model : $X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \qquad i = 1, \cdots, 10 \ ; \ j = 1, \cdots, 5$

where $\sum \alpha_i = \sum \beta_j = 0, X_{ij}$ denotes the estimated posterior mean for the ith data set with jth P value; α_i denotes the data set effect; β_j denotes the P effect; and $\epsilon_{ij}, i = 1, \cdots, 10; j = 1, \cdots, 5$ are independent error term having normal distribution with mean 0 and variance σ^2.

Our hypotheses are

$$H_0 \ : \ \beta_j = 0 \quad for\ all\ j \tag{4.1}$$

$$H_1 \ : \ \beta_j \neq 0 \quad for\ some\ j$$

The data set is as follows.

	No of iterations P				
Data set	10	20	30	50	100
1	.13660	.13843	.13855	.13723	.13295
2	.09013	.08949	.08967	.08883	.08994
3	.08796	.09021	.08942	.08927	.08890
4	.12364	.12467	.12528	.12395	.12465
5	.11747	.11720	.11727	.11756	.11734
6	.11574	.10634	.10826	.11237	.10846
7	.11259	.11383	.11323	.11265	.11328
8	.09413	.09416	.09409	.09471	.09404
9	.09184	.09122	.09115	.09071	.09070
10	.12175	.12149	.12168	.12130	.12164

Table 4.1: Estimated posterior means of simulated data

The results of the ANOVA is

Source	Sum of Square	df	MS	F - ratio
Treatment	SSTrt = .00000527	4	MST = 0.00000132	F = 0.58986044
Block	SSB = .01281011	9	MSB = 0.00142335	
Error	SSE =.00008042	36	MSE = 0.00000223	
Total	SST =.01289581			

Table 4.2: The ANOVA Table

We find that the p-value is 0.6722. Thus we cannot reject the null hypothesis at level of significance 0.05. That means the treatments have approximately the same effect. In our case, to lower the burden of the simulation, we choose $P = 10$.

4.2 Comparison between $\hat{R}_{bay}$ and $\hat{R}_{adjMLE}$

In this section, we compare the Bayesian estimator, $\hat{R}_{bay}$ with the independent one, $\hat{R}_{adjMLE}$ in terms of square error. The simulation study will be divided into three parts. They are (i) N is known, (ii) N is unknown and (iii) boundary problem. In all cases the initial setting for creating the observed profiles are the same.

The procedure is as follows. We adapt the assumption used in Chapter 3. The value of the unknown parameters N, and R will be fixed in simulating the observed profiles. Hence, we can simulate all these N spheres in the specimen by using the algorithm as described in Section 3.2.

After we have simulated all the spheres, we need to create the data set. We do it by fixing the section plane in the middle of the specimen. Next, we use

the algorithm in Section 3.3 to find the Bayesian estimator, $\hat{R}_{bay}$, and then to study the improvement on the estimator of R after imposing the dependence assumption.

In order to ensure that the adjusted maximum likelihood method of R can be used, we require that the number of the observed profiles is greater than or equal to one. However, for the case with dependent condition, even we cannot observe anything from the section plane, we can still use our algorithm to simulate R provided that N is known. In that case, the maximum radius of the observed profiles can no longer be used as our starting value of R.

Moreover, we want to study the effect of the intensity to our estimate. We roughly divide the intensity level into three classes. Recall that, through our discussion, we adapt V_v as a measure of the intensity. Thus we can change the intensity by varying the number of spheres.

In the simulation, we assume that the section plane is fixed in the middle of the specimen. This assumption is reasonable if we look at it from the view of boundary problem. Imagine that there is a substance with a huge volumn, and the specimen is cut from it. Obviously, the spheres inside the specimen will be affected by the outside ones. In order to consider this effect, we enlarge the specimen to a certain extent. Hence, we can enlarge the specimen such that the position of the section plane is in the middle. Therefore, it is not important where the position of the section plane is. On the contrary, we have benefit from this choice because more profiles can be obtained if the section plane is in the middle than that in the margin.

In the next three subsections, we will investigate the effect of the known or unknown number of spheres in the specimen and the boundary effect.

4.2.1 The number of spheres is known

In this subsection, we investigate the effect of the known number of spheres on the estimator of R. In order to study the effect of varied intensity, we divide the simulation into three parts, from low to high intensity. The meaning of low intensity and high intensity is just roughly differentiated. The highest intensity means the maximum number of the spheres that can be placed in the specimen without overlapping times the proportion of the sphere's volumn in the specimen and then divided by the volumn of the specimen. That is, high intensity $= \frac{\sum_{i=1}^{M} \frac{4\pi R^3 w_i}{3}}{V}$, where V is the volumn of the specimen;w_i is the proportion of the sphere i lying in the specimen, for example, if whole sphere is in the specimen, then $w_i = 1$; R is the true radius; and M is the maximum number of the spheres that can be placed in the specimen. An intuitive way to find M is to arrange the spheres in the specimen like the lattice. That is, $M = \lfloor \frac{l}{2R} + 1 \rfloor \lfloor \frac{w}{2R} + 1 \rfloor \lfloor \frac{h}{2R} + 1 \rfloor$, where $\lfloor x \rfloor$ is the greatest integer which is smaller than or equal to x. However, it is not the most closely packed form. This form can be achieved if the centres of the adjacent spheres are arranged to form an equilateral triangle. In our case, we just need to have an idea of how many spheres in the specimen representing the low, medium and high intensity, so we can use the former, the simpler one.

For the low, medium and high intensity, we simulate $N = 10, 15, 18$ spheres respectively with $R = 0.2$ in the unit cubic specimen by using the algorithm in

Section 3.2. Then we follow the algorithm in Section 3.3 and get the posterior mean of R, $\hat{R}_{bay}$ as an estimate of the constant R. We do it for 150 different sets of data. For the comparison, we calculate the square error for each data set i and denotes them as $SE1_i$ and $SE2_i$, $i = 1, \cdots, 150$, for the independence and dependence case respectively. That is, for $i = 1, \cdots, 150$,

$$\begin{aligned} SE1_i &= (\hat{R}_{adjMLE,i} - R)^2 \\ SE2_i &= (\hat{R}_{bay,i} - R)^2, \end{aligned} \tag{4.2}$$

where $\hat{R}_{bay,i}$ and $\hat{R}_{adjMLE,i}$ are the Bayesian estimator, independent estimator for data set i respectively, and R is the true radius.

Afterward, we compute $log_{10}(ratio_i) = log_{10}(SE1_i/SE2_i)$ to see which one is better. The value of $log_{10}(ratio_i)$, which is greater than zero represents that $\hat{R}_{bay,i}$ is better than $\hat{R}_{adjMLE,i}$, in terms of square error. The results are in the following figures from the low intensity to the high intensity. Besides, we calculate the proportion of $\hat{R}_{bay}$ that is better than $\hat{R}_{adjMLE}$ in Table 4.3.

Intensity Level	Low	Medium	High
Percentage that $\hat{R}_{bay}$ is better than $\hat{R}_{adjMLE}$	66	76	81

Table 4.3: Percentage that $\hat{R}_{bay}$ is better than $\hat{R}_{adjMLE}$ (Known N)

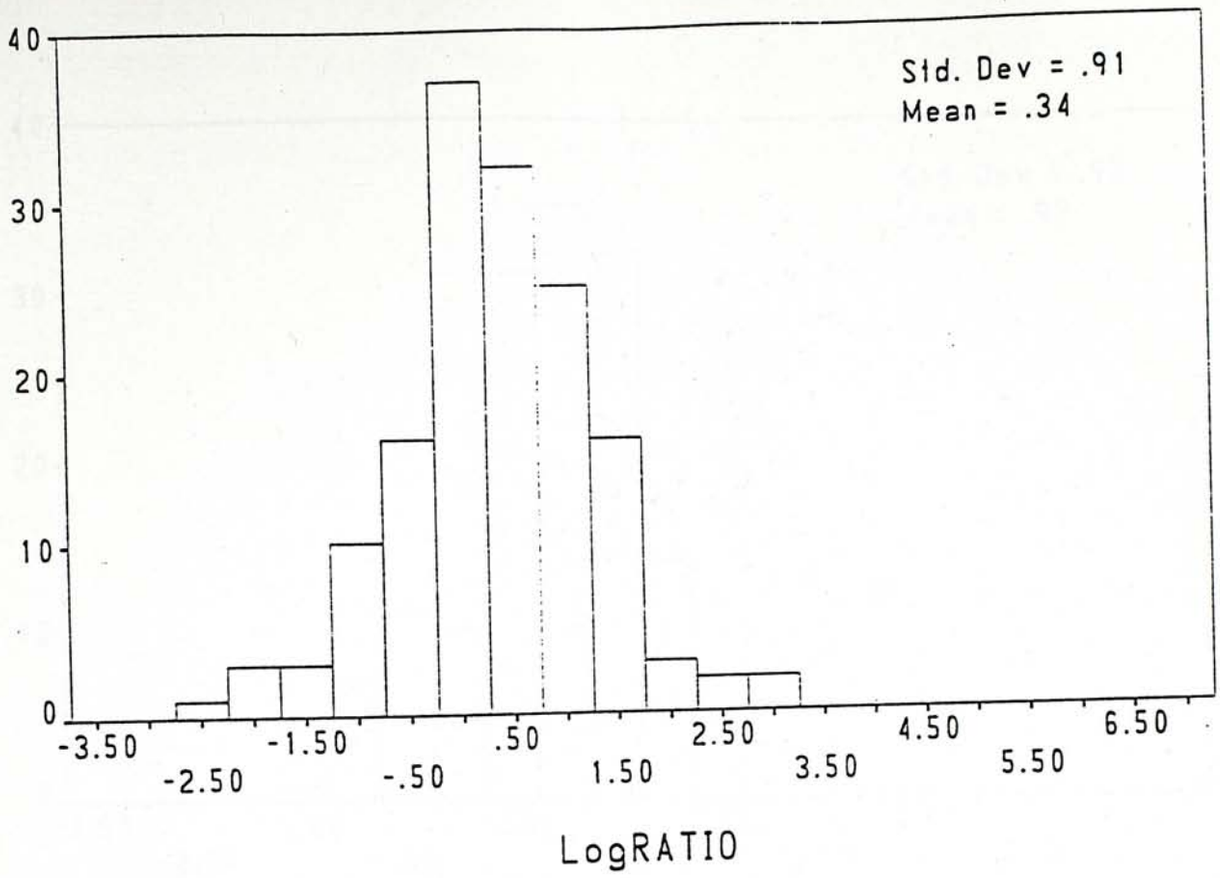

Figure 4.1 The $log_{10}(ratio)$ with the low intensity

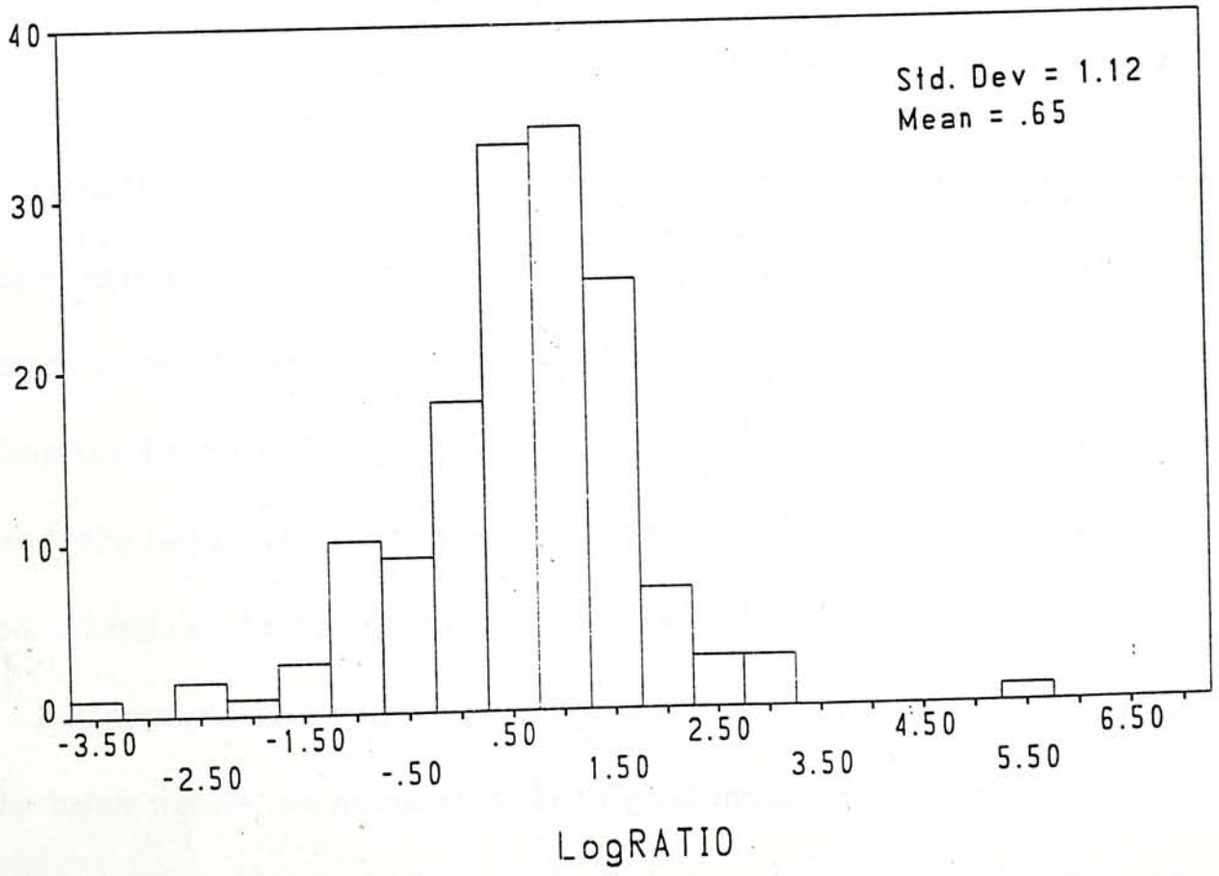

Figure 4.2 The $log_{10}(ratio)$ with the medium intensity

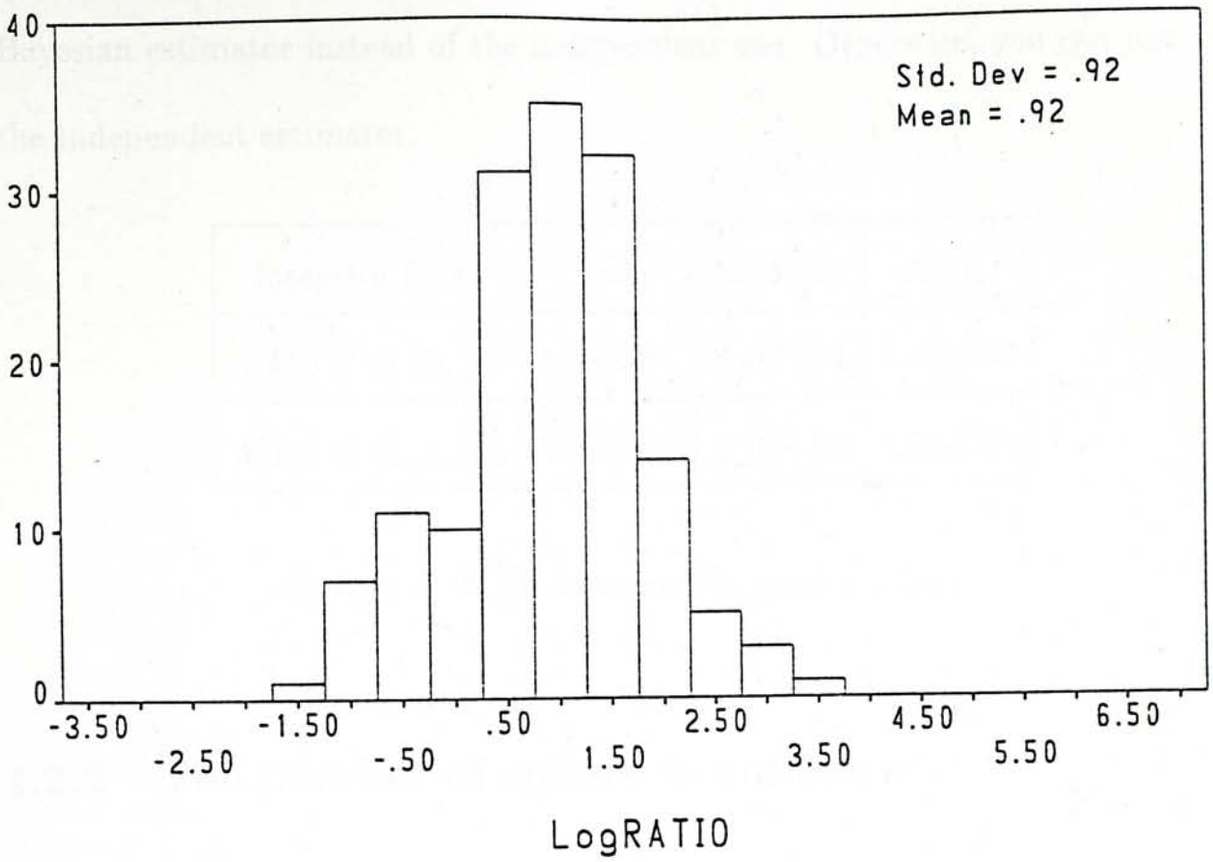

Figure 4.3 The $log_{10}(ratio)$ with the high intensity

From the above figures, we find that when the intensity level increases, the improvement of $\hat{R}_{bay}$ increases too. It seems reasonable because when the intensity level increases, the number of spheres cut by the section plane increases. Thus the Bayesian estimator $\hat{R}_{bay}$ will become more accurate. On the other hand, the larger the sample size, the more accurate the independent estimator too. Therefore, the overall proportion will not be so high.

Furthermore, we list the MSE of the above two estimators in Table 4.4. From the below figures, we notice that the original independent estimator is already quite good in terms of MSE, while the Bayesian estimator also contributes some improvements in the amount of MSE. As the computation time is not expen-

sive nowadays, if you want estimator with high precision, you should use the Bayesian estimator instead of the independent one. Otherwise, you can just use the independent estimator.

Intensity Level	Low	Medium	High
MSE of $\hat{R}_{bay}$	1.18E-04	3.46E-05	7.19E-06
MSE of $\hat{R}_{adjMLE}$	2.74E-03	5.12E-05	1.56E-05

Table 4.4: MSE between $\hat{R}_{bay}$ and $\hat{R}_{adjMLE}$

4.2.2 The number of sphere is unknown

We change to the condition that N is unknown. The estimator in (3.11) is used to estimate N. The setting of the simulation is the same as the Section 4.2.1. We also divide the simulation into three parts from the low intensity to the high intensity. The results are summarized in the following figures and tables.

Intensity Level	Low	Medium	High
Percentage that $\hat{R}_{bay}$ is better than $\hat{R}_{adjMLE}$	52	72	83

Table 4.5: Percentage that $\hat{R}_{bay}$ is better than $\hat{R}_{adjMLE}$ (Unknown N)

From Table 4.5, we see the trend of the improvement of $\hat{R}_{bay}$ as the intensity level increases. However, the amount of improvement is smaller than that when N is known, especially in the low intensity. An obvious reason is that the information of N plays an important role in the estimation of R.

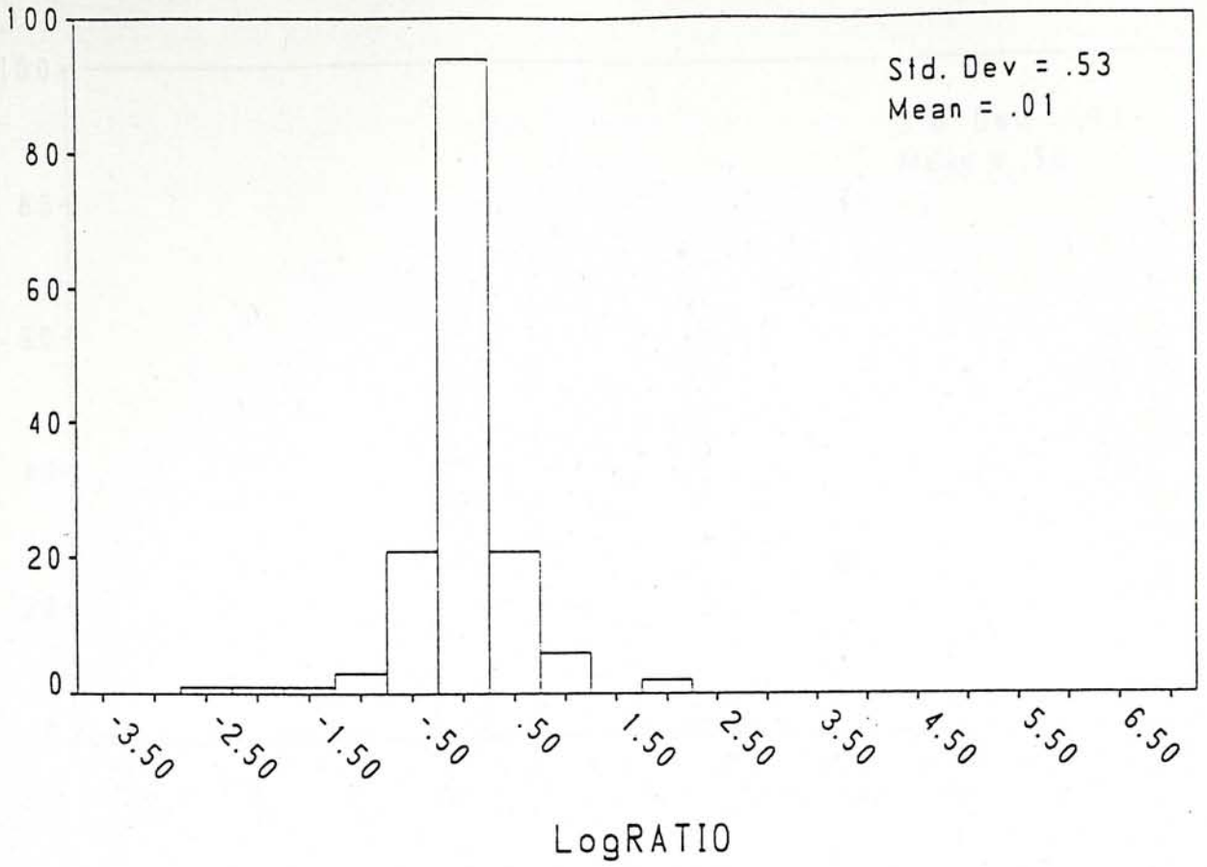

Figure 4.4 The $log_{10}(ratio)$ with the low intensity

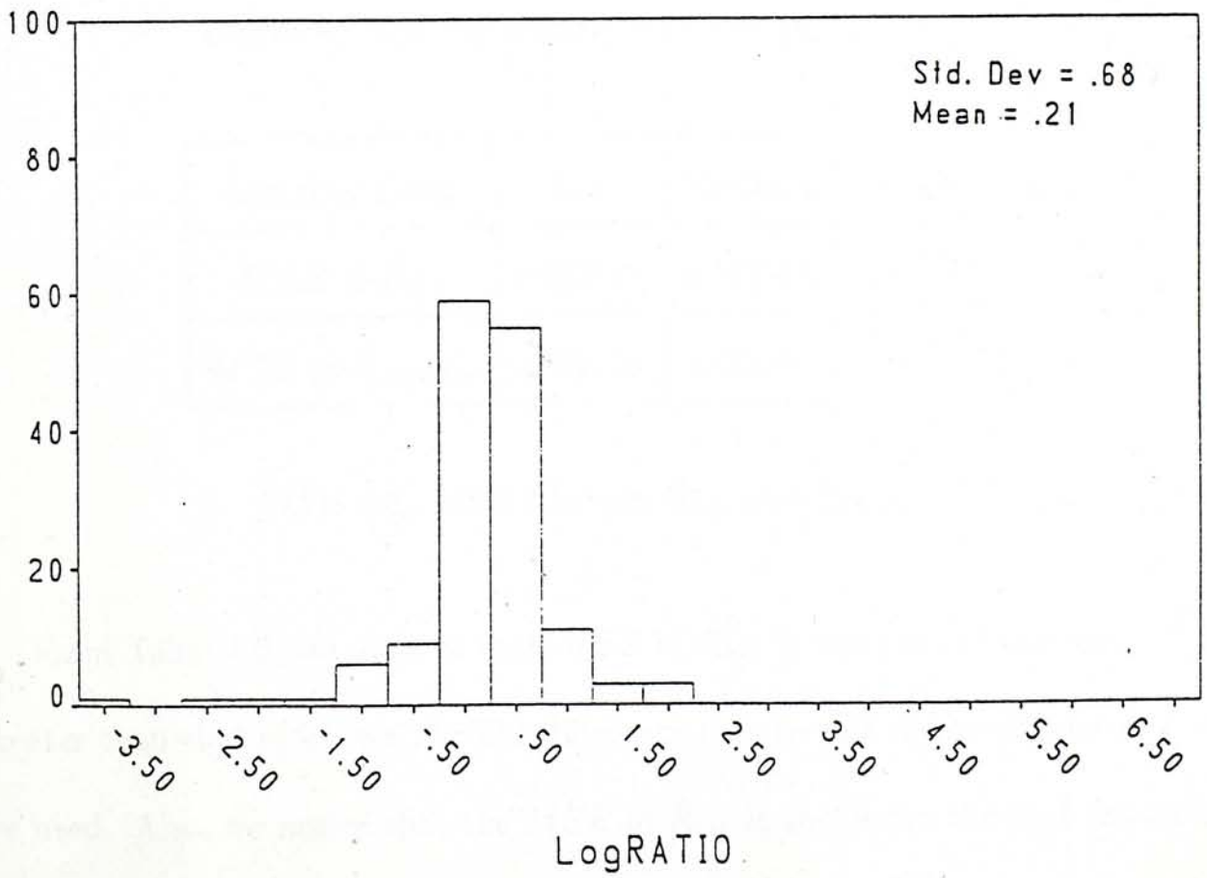

Figure 4.5 The $log_{10}(ratio)$ with the medium intensity

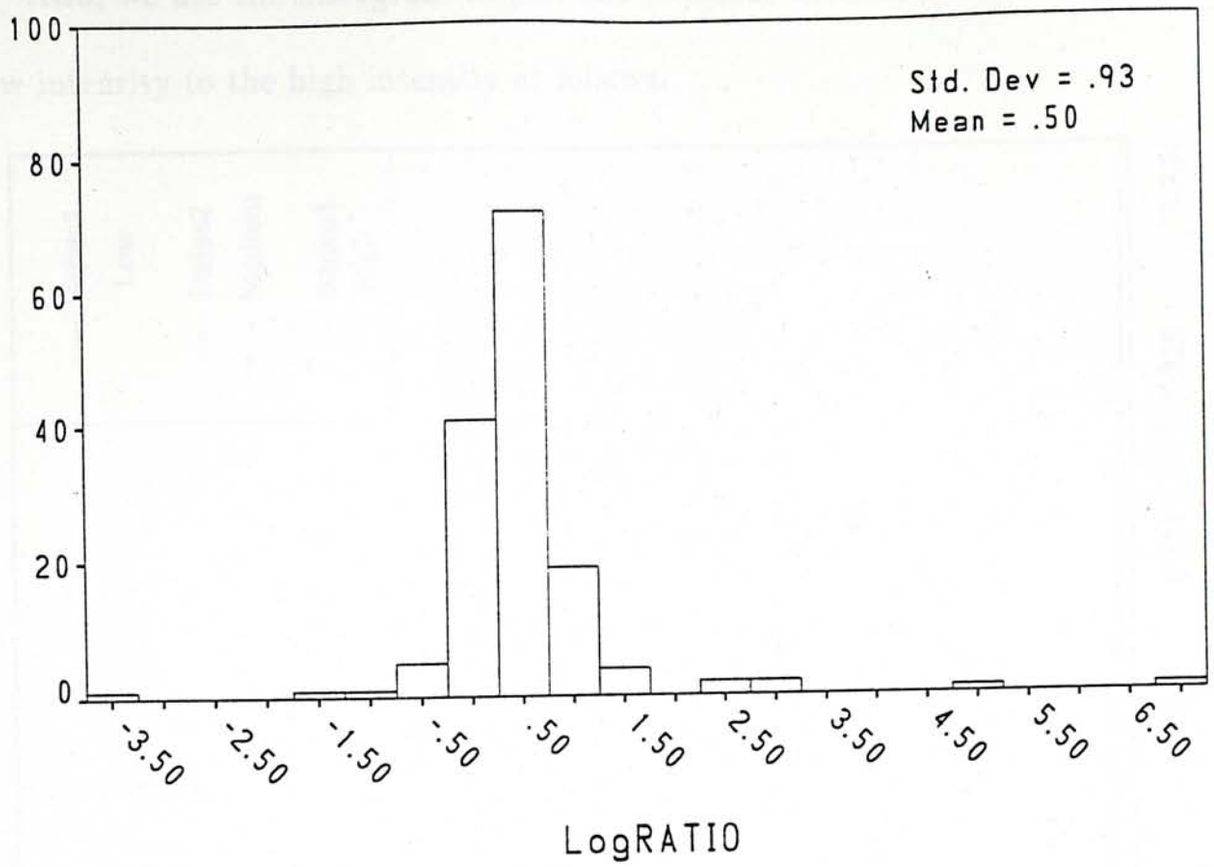

Figure 4.6 The $log_{10}(ratio)$ with the high intensity

Intensity Level	Low	Medium	High
MSE of $\hat{R}_{bay}$	4.38E-04	3.59E-05	1.31E-05
MSE of $\hat{R}_{adjMLE}$	2.99E-04	3.81E-05	1.75E-05

Table 4.6: MSE between $\hat{R}_{bay}$ and $\hat{R}_{adjMLE}$

From Table 4.6, we observe that MSE of $\hat{R}_{bay}$ in the case of unknown N is greater than that of known N. The difference may be due to the estimator of N we used. Also, we notice that the MSE of $\hat{R}_{bay}$ is smaller in the high intensity case. It is consistent with the case of N is known. For the independent estimator, $\hat{R}_{adjMLE}$, the MSEs are quite similar to the case of known N.

Also, we use the histogram to plot the posterior distributions of R from the low intensity to the high intensity as follows.

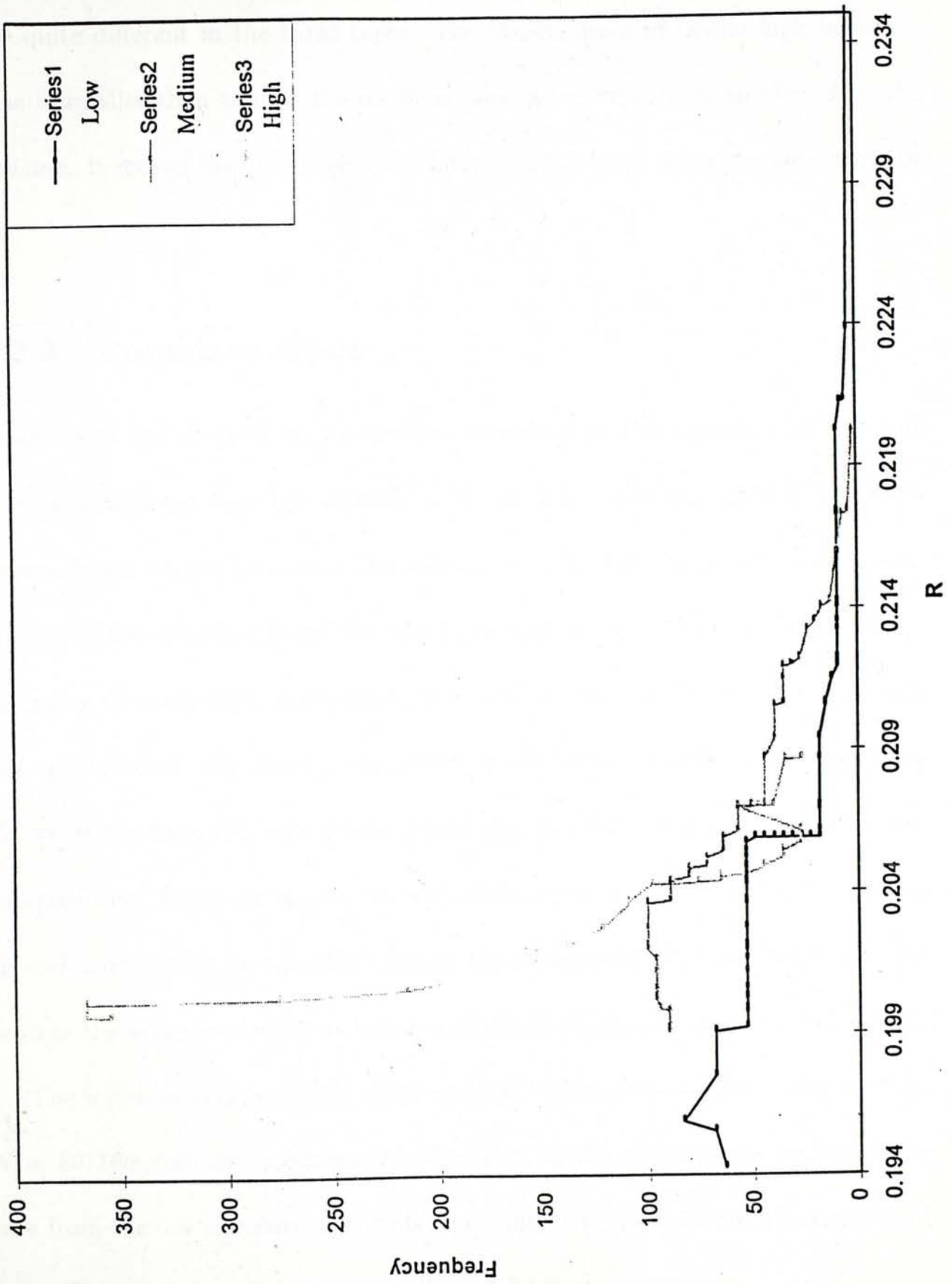

Figure 4.7 The posterior distribution of R with intensity from low to high

From Figure 4.7, series 1, 2, 3 represent the posterior distributions from the low to high intensity respectively. We notice that the shape of the distributions are quite different in the three cases. The dispersion of R in the high intensity case is smaller than that in the medium case, which in term is smaller than the low one. It means that the higher the intensity, the more accurate our estimator is.

4.2.3 Boundary effect

Apart from the effect of N, we are also interested in the boundary effect. The setting is different from the sections 4.2.1 and 4.2.2. The raw data is simulated from a larger region instead of the origin one such that it can take into consideration of the influence from the outside spheres to the inside one. We do it by enlarging the unit cubic specimen to a larger one with the length $Ext = 2$ times the original one. We also fix the section plane in the middle of the specimen. To create the data set, we concentrate on the unit cubic specimen instead of the enlarged one. Next, we enlarge the unit cubic specimen by $2 \times Amt \times \hat{R}_{adjMLE}$ on each side so that we can also consider the influence of the outside spheres. We process the same procedure as before and use the posterior mean to estimate R.

The initial setting is slightly different from the previous sections. We simulate $N = 80, 160$ and, 240 spheres with $R = 0.12$ in the specimen to represent the case from the low intensity to the high intensity. We do it for 50 different sets of data. The results are listed in the following figures and tables.

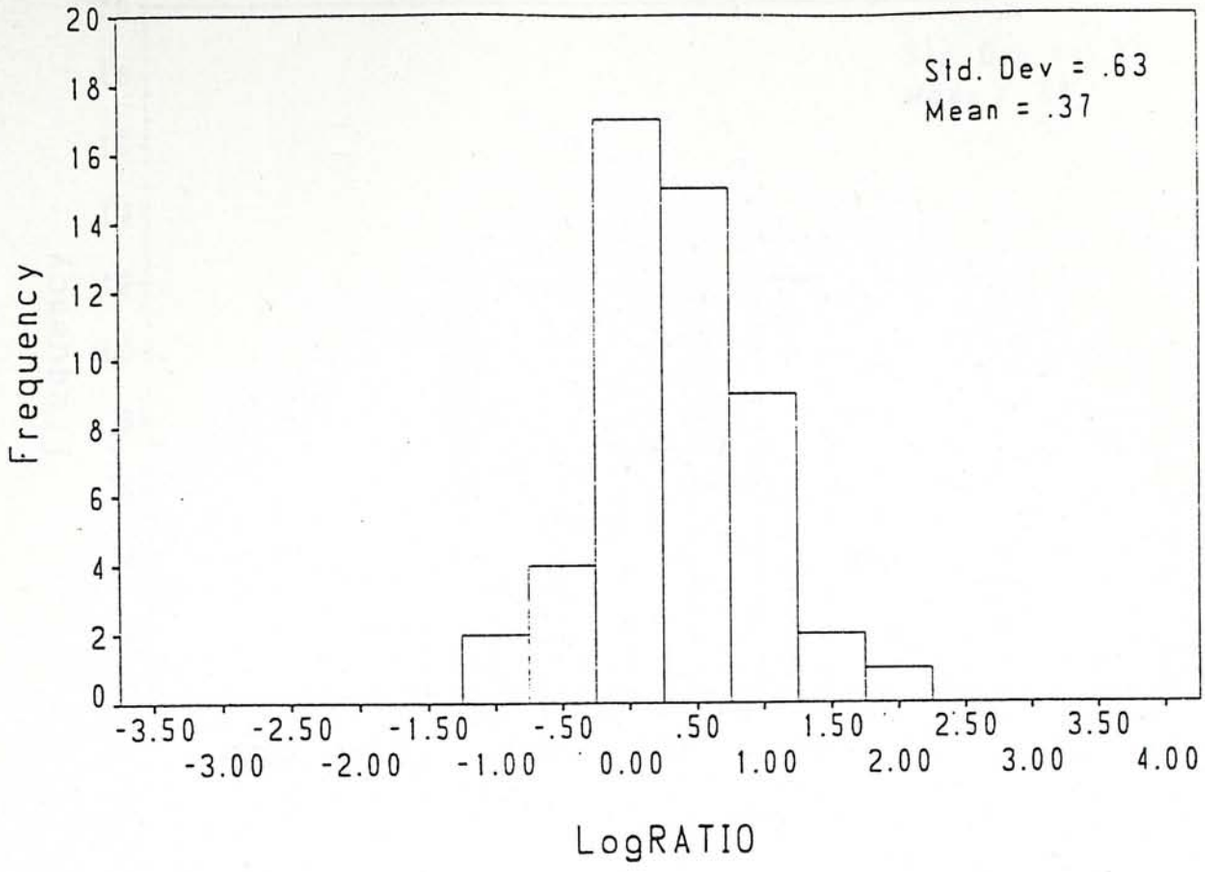

Figure 4.8 The $log_{10}(ratio)$ in the low intensity with $AMT = 1$

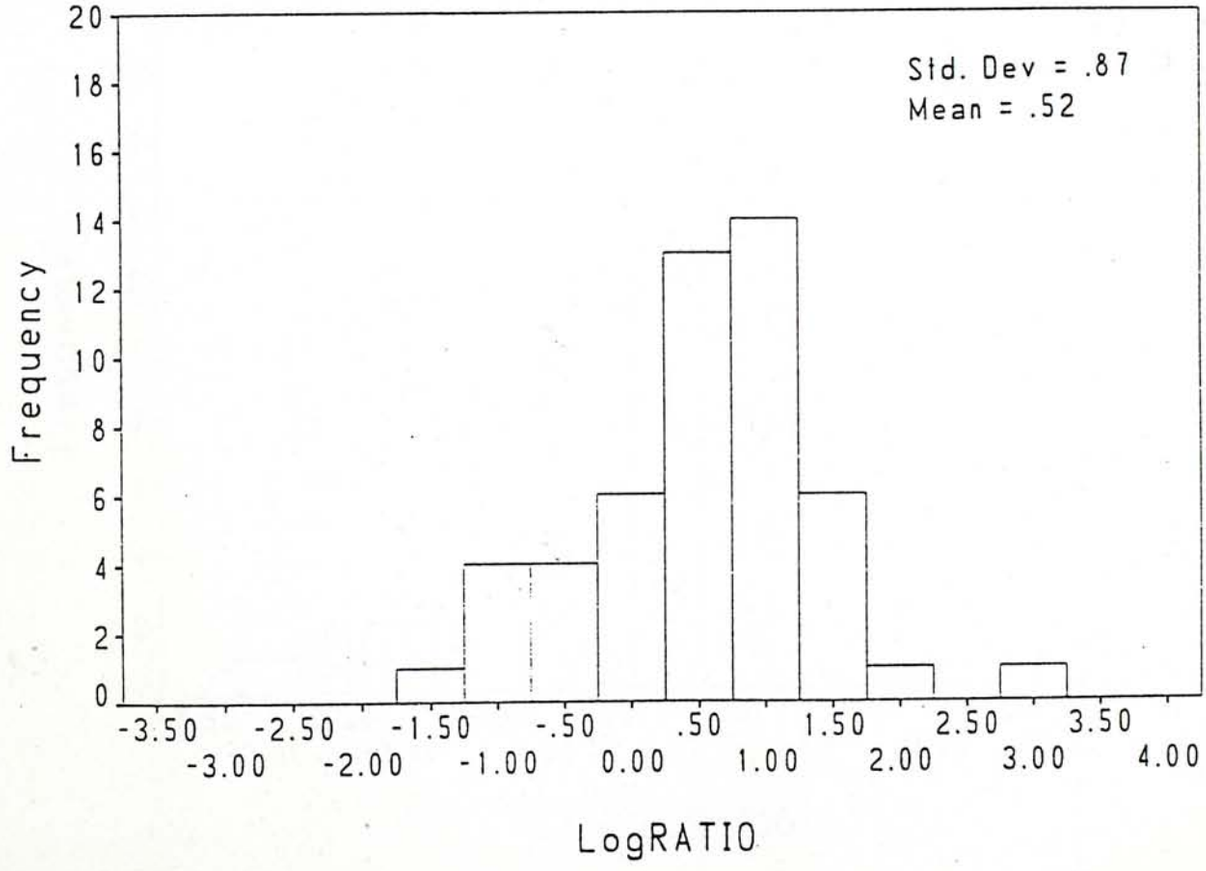

Figure 4.9 The $log_{10}(ratio)$ in the low intensity with $AMT = 2$

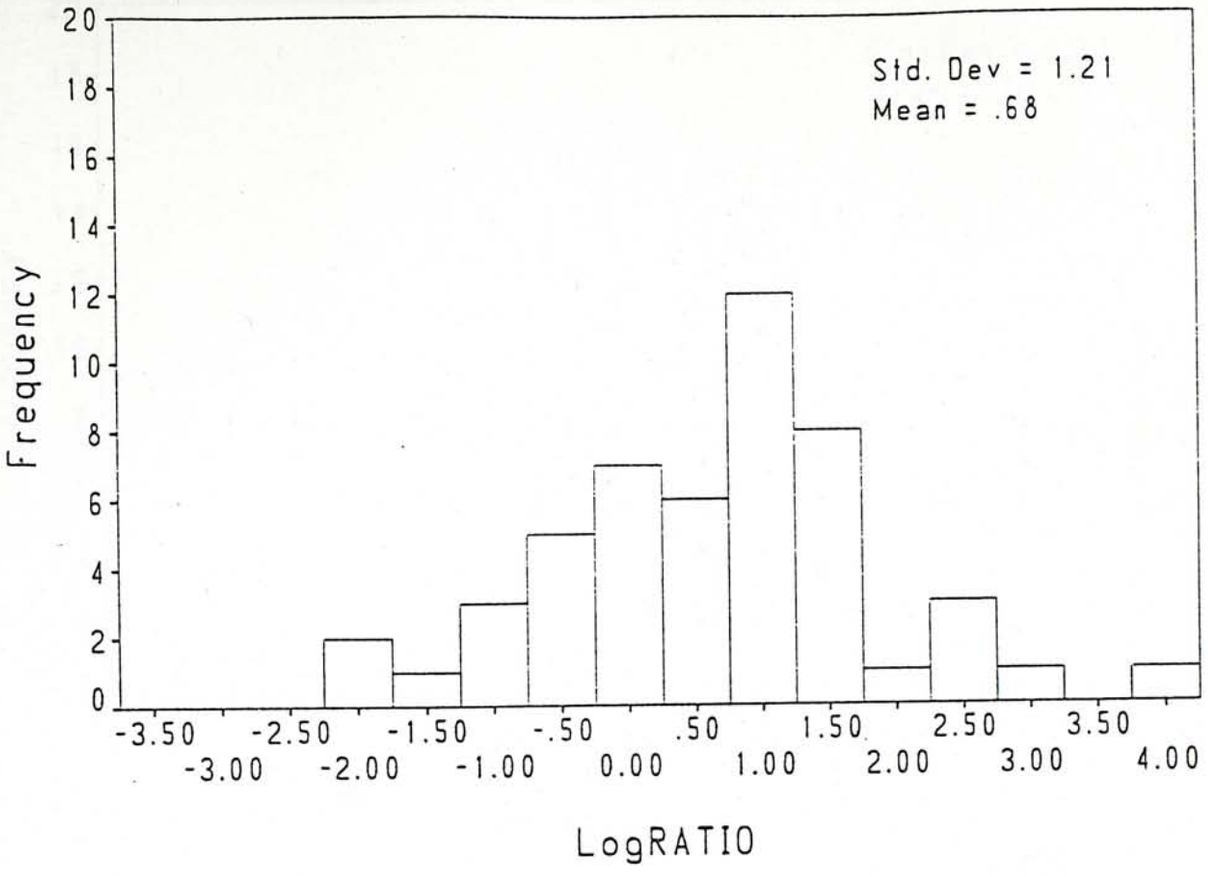

Figure 4.10 The $log_{10}(ratio)$ in the low intensity with $AMT = 3$

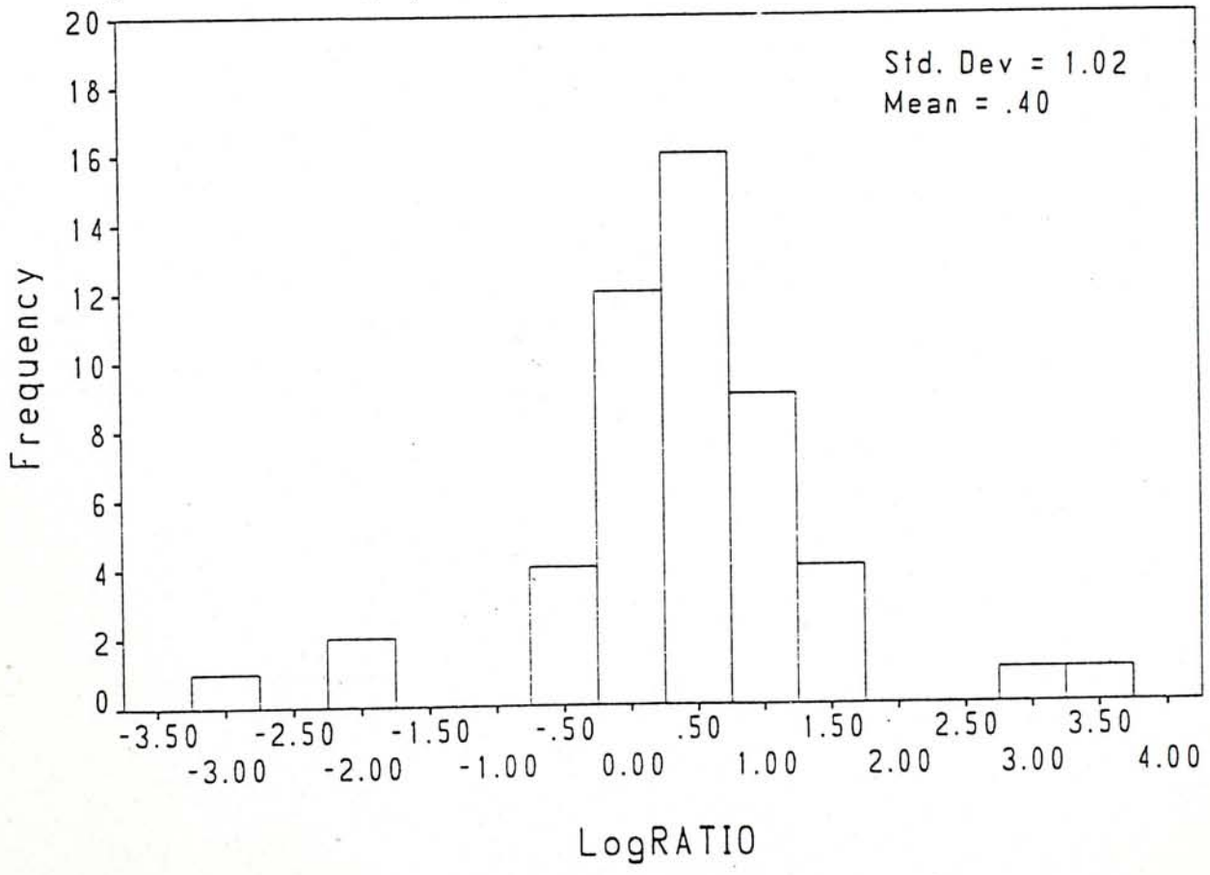

Figure 4.11 The $log_{10}(ratio)$ in the medium intensity with $AMT = 1$

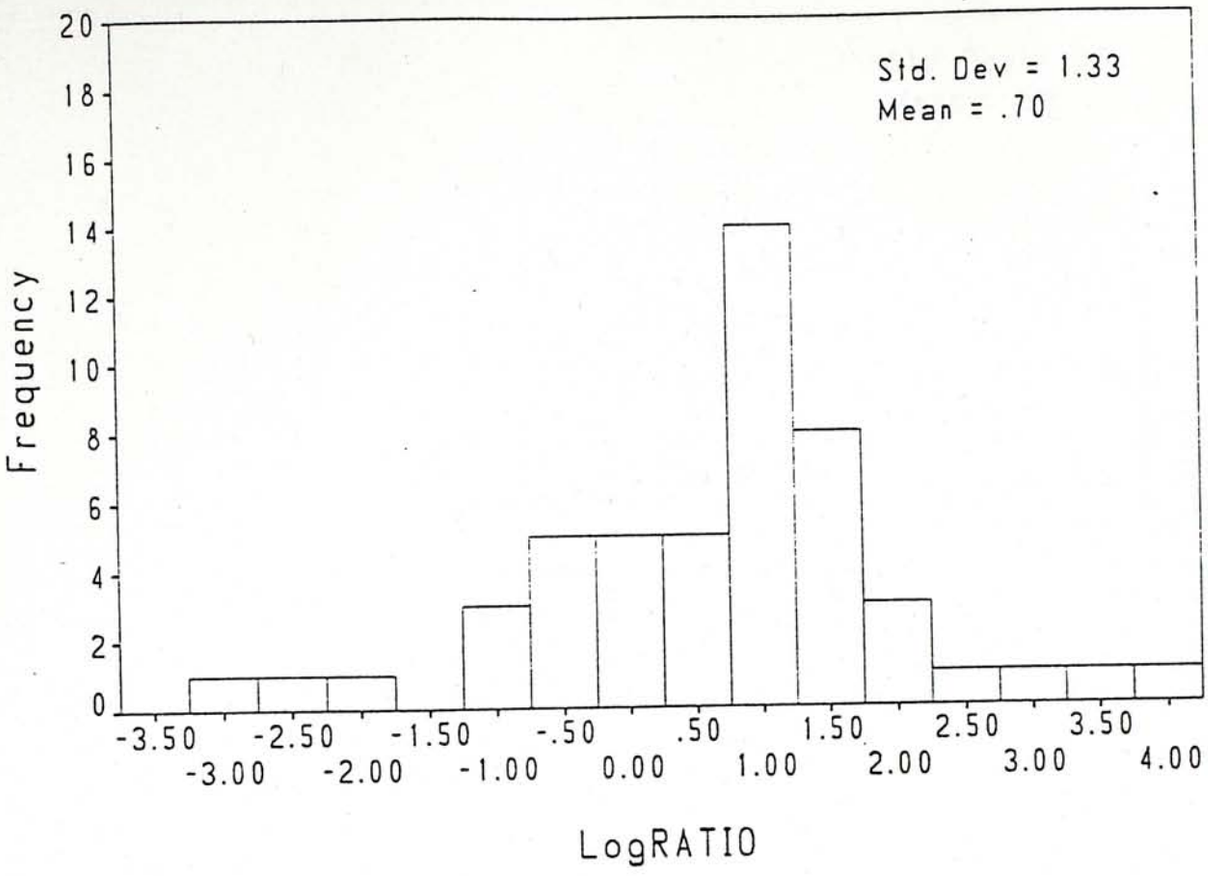

Figure 4.12 The $log_{10}(ratio)$ in the medium intensity with $AMT = 2$

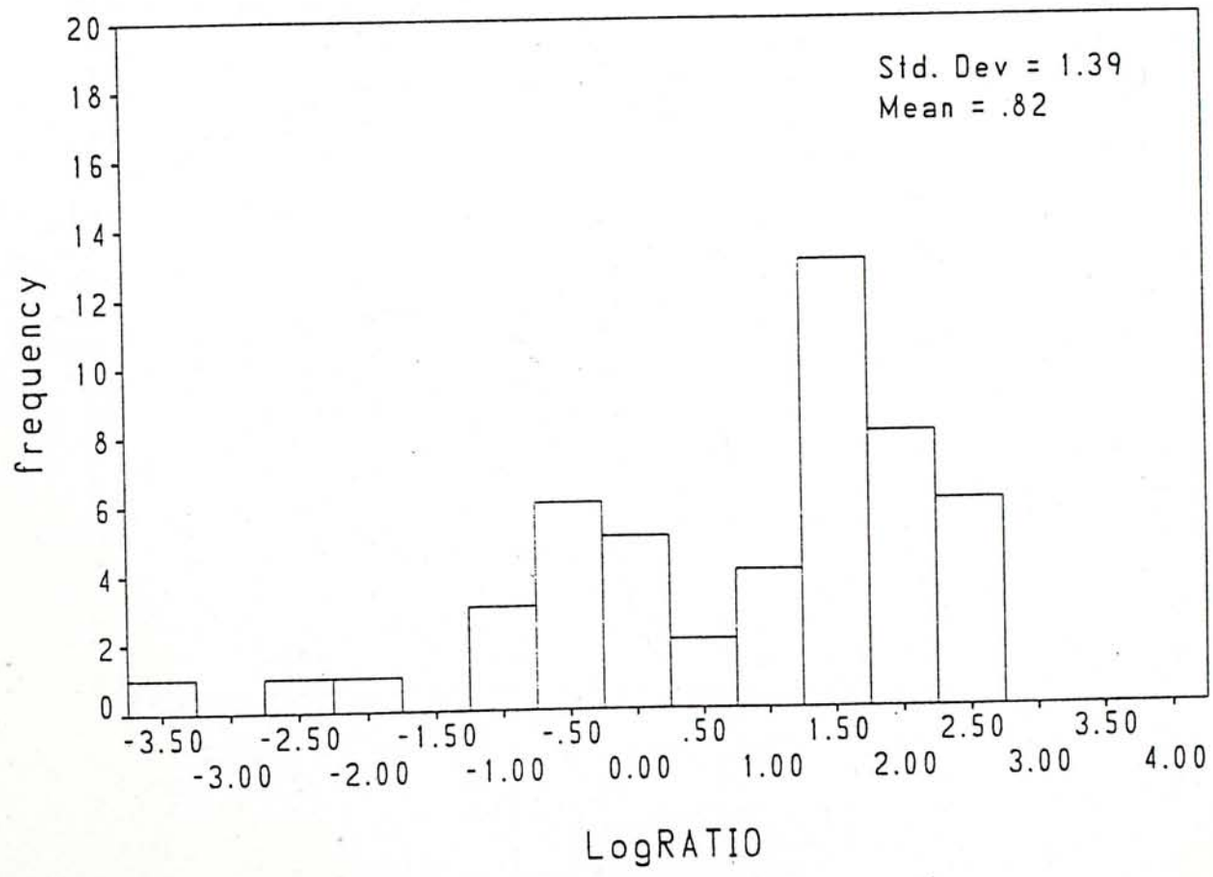

Figure 4.13 The $log_{10}(ratio)$ in the medium intensity with $AMT = 3$

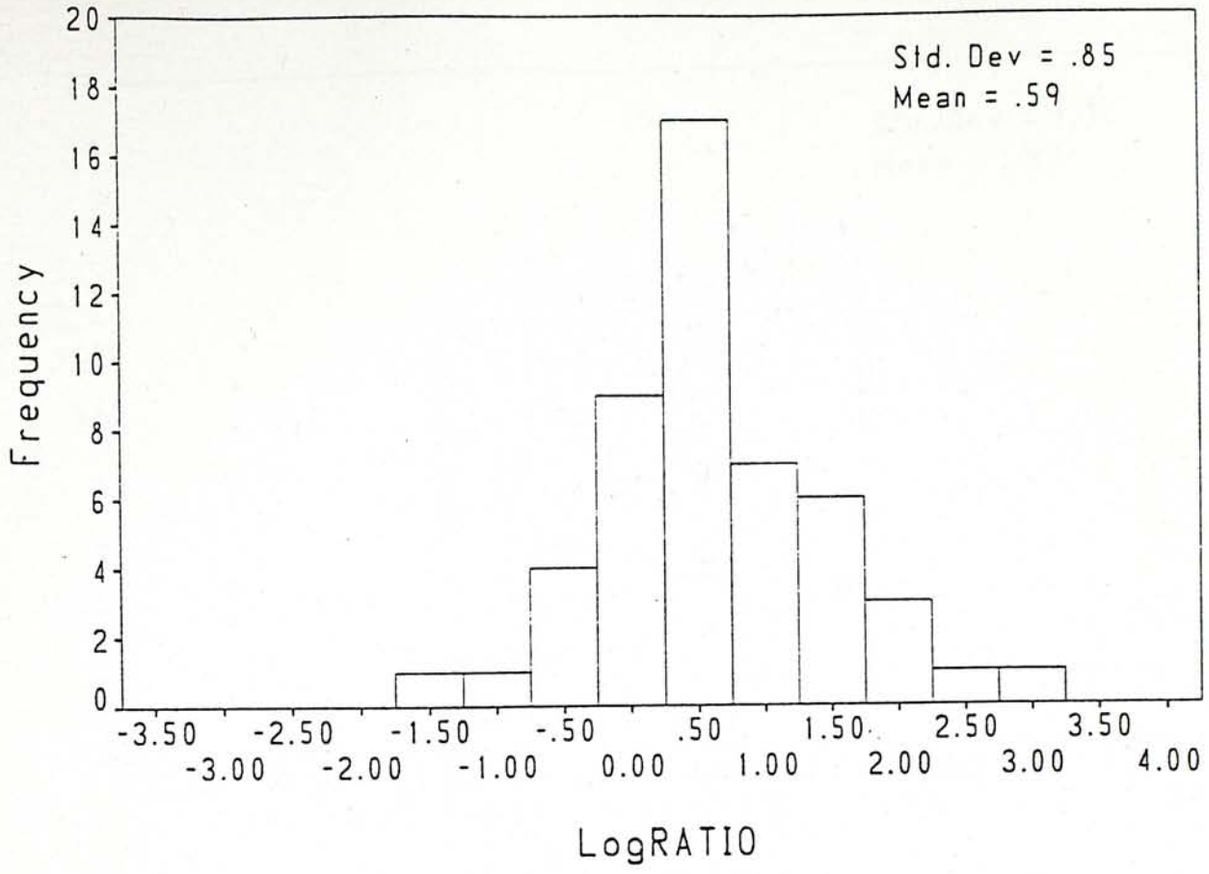

Figure 4.14 The $log_{10}(ratio)$ in the high intensity with $AMT = 1$

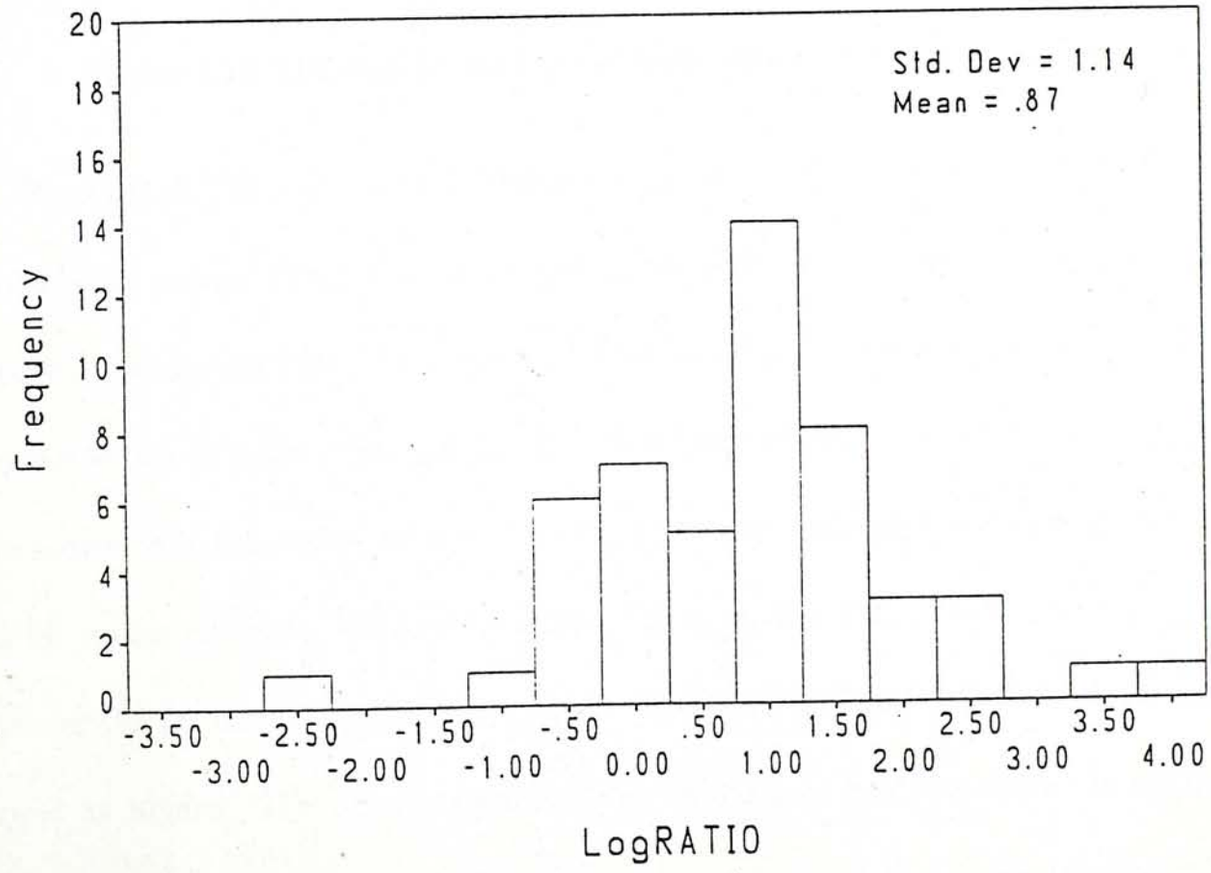

Figure 4.15 The $log_{10}(ratio)$ in the high intensity with $AMT = 2$

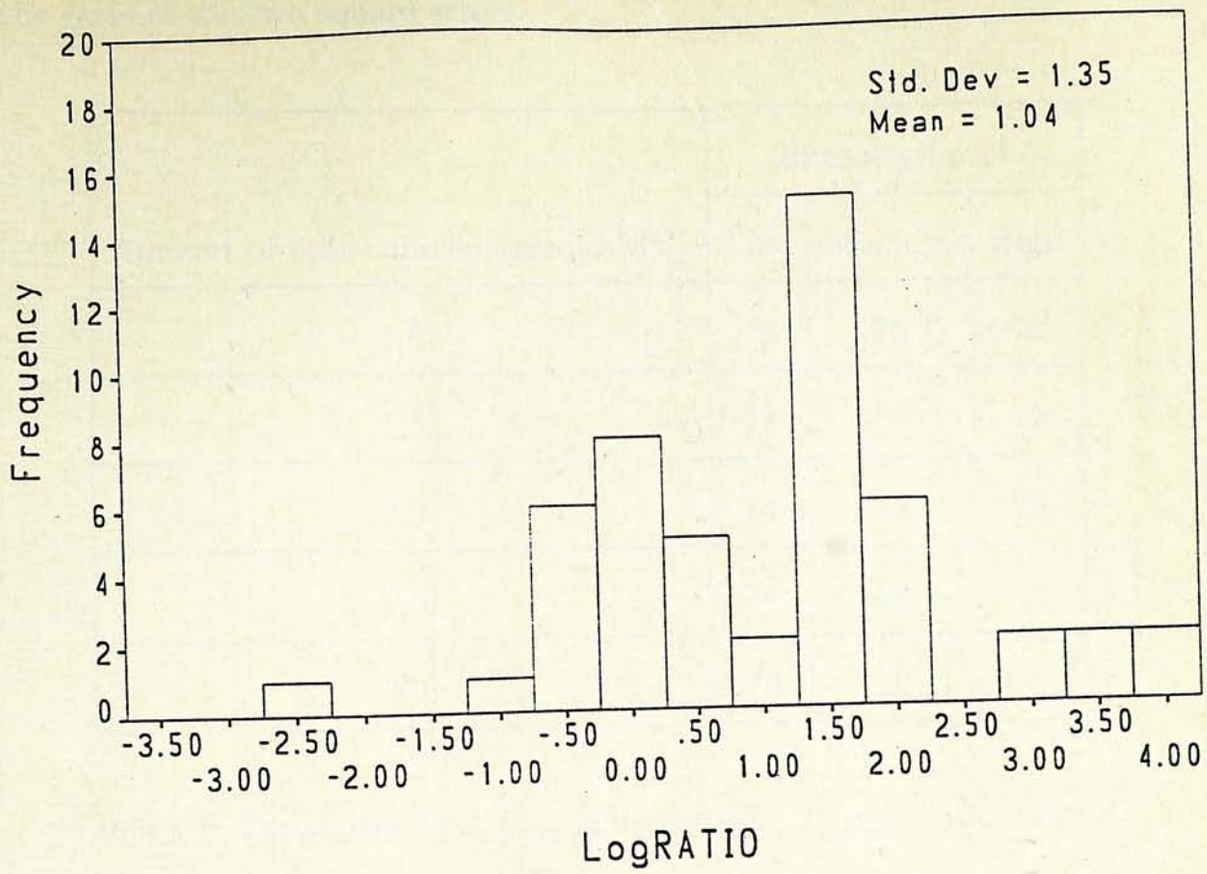

Figure 4.16 The $log_{10}(ratio)$ in the high intensity with $AMT = 3$

Recall that the x axis of the histogram represents the logarithm ratio of the two square errors. From the above figures, we focus on the mean value on each figure for each intensity level. We find that as AMT increases, the mean value increases too. On the other hand, conditional on the same amount of extension, we notice that the mean value increases as intensity increases. The result seems to be consistent with the previous one. That is, the higher the intensity, the greater the improvement. However, from Table 4.7, we find that there is not any trend as before. The percentages seem to be more or less the same. It means that the amount of extension of boundary from $Amt = 1$ to 3 does not improve the proportion on our estimate very much. Nevertheless, it enhances the amount

of the ratio of the two square errors.

Amount of unit cube enlarged(AMT)	Intensity Level		
	Low	Medium	High
1.0	70	70	72
1.5	74	72	72
2.0	74	70	70
2.5	74	70	70
3.0	70	70	68

Table 4.7: Percentage that $\hat{R}_{bay}$ is better than $\hat{R}_{adjMLE}$ (Boundary)

Besides, we list the $MSEs$ of the two estimators with the different levels of AMT in Table 4.8, and 4.9.

Amount of unit cube enlarged(AMT)	Intensity Level		
	Low	Medium	High
1.0	3.36E-04	1.05E-04	4.19E-05
1.5	3.46E-04	9.58E-05	4.29E-05
2.0	3.15E-04	1.01E-04	4.98E-05
2.5	2.90E-04	9.65E-05	4.97E-05
3.0	2.84E-04	9.82E-05	5.33E-05

Table 4.8: MSE of $\hat{R}_{bay}$ with different level of AMT value

Amount of unit cube enlarged(AMT)	Intensity Level		
	Low	Medium	High
1.0	4.34E-04	1.17E-04	4.59E-05
1.5	4.34E-04	1.17E-04	4.59E-05
2.0	4.34E-04	1.17E-04	4.59E-05
2.5	4.34E-04	1.17E-04	4.59E-05
3.0	4.34E-04	1.17E-04	4.59E-05

Table 4.9: MSE of $\hat{R}_{adjMLE}$ with different level of AMT value

From Table 4.8, if we focus on each column, i.e. conditional on the same intensity, the values of MSEs of $\hat{R}_{bay}$ are quite similar no matter what values of AMT are. If we look at the value on each row, i.e. conditional on the value of AMT, the MSEs seem to be same as before. That is, the smaller MSE in the high intensity case.

From Table 4.9, the MSEs of $\hat{R}_{adjMLE}$ are quite consistent with the previous two cases, N is known and N is unknown. In summary, the extension of the boundary, from $AMT = 1$ to 3, can give us some improvements on our estimate of R.

Chapter 5

Extension and Conclusion

As mentioned in the introduction, the motivation of this thesis is due to the dependence among the spheres. However, it does not imply that the situation discussed in Chapter 2 is not possible to occur. For example, if the spheres are represented by the gas bubbles, they will be independently distributed among the specimen. The problem of influence on each other will not happen in this case because they can cross through each other arbitrarily. Therefore, it is still worth to study the case of independence.

5.1 Extension on the simulation algorithm

Although we have discussed the simulation algorithm in detail in the previous chapters, there are still some minor problems which have to be solved for the extension to the general case. We will illustrate as follows.

Relation between R and U, C

In Chapter 3, we assume that the distribution of $R \mid (U, C)$ is uniform. If the conditional distribution of R is not uniform, it is more difficult to achieve the posterior distribution of R. Instead of finding $\Pi(R \mid S_{com})$ directly by using the two methods in Section 3.6, we approximate $\Pi(R \mid X_{obs})$ by the following steps.

1. After we iterate $P(= 10)$ steps to get R, we stop and based on the final observations of $\vec{u}_P$ and c_P, we further simulate $T(= 30)$ more R's and then plot the histogram for these R values. This histogram can be used to approximate the posterior distribution of R with complete s_i, $\Pi(R \mid S_{com})$, because the results from Gibbs Sampler shows that these R's are effectively the points from the marginal density of R when P is large.

2. We iterate $Q(= 20)$ more times and get Q number of the posterior distributions, then averaging these posterior distributions and we can obtain the true posterior distribution, $\Pi(R \mid X_{obs})$ as before.

The unknown number of N

In Section 3.6, we propose a reasonable and simple estimator of N. Apart from this estimator, we suggest another estimator of N. The idea of this method comes from stereology. Before presenting the method, we define some terms and extract some useful results from stereology. Define A_a as the expected areal density of profiles on the section, that is the ratio of the sum of profiles areas to the sum of section areas, and V_v as the volume density which is the total volumn of spheres divided by the volumn of specimen.

The Principle of Delesse

Suppose we place a unit cubic specimen into the $(x, y, z)-$ coordinate system and the section plane, which is parallel to the x, y plane, cut the specimen randomly. That is, the distance of the section plane from the origin follows $U(0, h)$, where h is the height of the specimen. Then, it can be shown that $A_a = V_v$. It gives us an estimator of V_v and it can be proved that A_a, is an unbiased estimator of V_v under the above condition.

To estimate N, we first calculate A_a. Then, we can estimate V_v by A_a, and hence N can be estimated from the following equation,

$$V_v = \frac{4\pi \hat{R}^3 \hat{N}}{3V}, \tag{5.1}$$

where V is the volume of the specimen. However, we also face the same problem as before. What is our estimate of R? One of the possible choice is $\hat{R}_{adjMLE}$. Besides, we have a suggestion to combine the above method and the one discussed in Section 3.6 to estimate the two unknown parameters N and R. We can divide the simulation into two stages. In the first stage, we solve the above two equations and get the values of N and R. The result is as follows.

$$\hat{N} = \sqrt{\frac{\pi M^3 V^2}{6 V_v A^3}} \tag{5.2}$$

$$\hat{R} = \sqrt{\frac{3AV_v}{2M\pi}} \tag{5.3}$$

Clearly, the estimated value of N may not be an integer. We need to round-off it using randomizing technique. For example, if we get $\hat{N} = 10.3$, we can decide the

randomized technique such that the probability of getting $N = 11$ is equal to 0.3, while having 0.7 probability N equals to 10, and the mean value is equal to 10.3 which is the calculated value. The second stage is that, based on the calculated value of $\hat{N}$, we use the same algorithm as before to estimate the R. The starting value of R can be the estimate of R in (5.3). That is, the initial value of R is $max(R(1), \hat{R})$, where $R(1)$ is the maximum profiles radius. Moreover, we can use this simulated value of R to estimate V_v again.

The behaviour of this estimator depends on the estimator of A_a. To make sure that it is unbiased, the section plane should be random instead of being fixed to the middle.

5.2 Extension to the case of varied radius

From the previous chapters, we notice that under different assumptions, such as (i) N is known, (ii) N is unknown and (iii) the boundary problem. Our proposed algorithm will give us a better estimator, $\hat{R}_{bay}$ than the independent one $\hat{R}_{adjMLE}$, in terms of square error. However, we impose some conditions that are different from the stereologist's interest, such as we assume that the spheres have equal radius. As mentioned before, the condition of constant radius is adopted because of the simplicity in modelling the spatial distribution of the spheres. Now, we want to extend to the case of varied radius.

5.2.1 Modified Algorithm

The main procedure of the algorithm is quite similar to the one discussed in Section 3.3, and also we use the same notations, but now R is an $N \times 1$ random vector $(R_1, \cdots, R_N)'$, where R_k is the true radius for the sphere k, $k = 1, \cdots, N$. In this case, we assume that N is known.

[1] Conditional on the observed profiles, denoted as X_{obs}, we find the feasible regions of R_i for $i = 1, \cdots, N$.

For the seen sphere, $R_i \sim U(x_i, R_{u,i})$, $i = 1, \cdots, n$.

For the unseen sphere, $R_i \sim U(0, R_{u,i})$, $i = n+1, \cdots, N$.

where x_i is the profile radius of the i th seen sphere, $i = 1, \cdots, n$ and $R_{u,j}$ is the upper bound of the radius of the j th sphere, $j = 1, \cdots, N$.

The starting value of R for the seen and unseen sphere are x_i and 0 respectively. The upper bound of R_i, i.e. $R_{u,i}$ is difficult to be evaluated because it depends on the other $R_{u,j}$, $j \neq i$. However, it is not necessary for us to determine it explicitly. Thus we let

$$R_u = \max_{1 \leq j \leq N} R_{u,j}$$

and the determination of R_u is just discussed in section 3.4. Then, we use this R_u for the upper bound of all R_k in [6].

[2] Initialize the variable U.

[3] Deterministically locate the unseen sphere C.

[4] Repeat [4] - [7] P times to get one sequence of P values of R.

[5] Given R and U, simulate C.

[6] Given U and C, simulate R.

[7] Given C and R, simulate U.

[8] Basing on the final observations of $C = c_P$ and $U = \vec{u}_P$, we further simulate T more R-values and treat them as the true samples from the distribution.

[9] Calculate the posterior distribution of R by drawing the R values which simulated from [8].

[10] Repeat [1]-[9] Q times and then draw a histogram of the above simulated R values to approximate the true posterior distribution of $R \mid X_{obs}$.

5.2.2 An artificial example

We use an artificial example to demonstrate the above algorithm. We simulate $N = 10$ spheres with varied radius in a unit cubic specimen. The radii are 0.06, 0.09, 0.1, 0.1, 0.1, 0.11, 0.12, 0.12, 0.18, and 0.19. Their mean, say R_t and standard deviation are 0.117 and 0.03974 respectively. We fix the section plane in the middle to create the data set X_{obs} as before. Throughout the simulation, we assume that N is known to be 10 and P, Q, T are 10, 20, 50 respectively. Then, we apply the above algorithm to estimate the posterior distribution of R. The result is listed in the Figure 5.1.

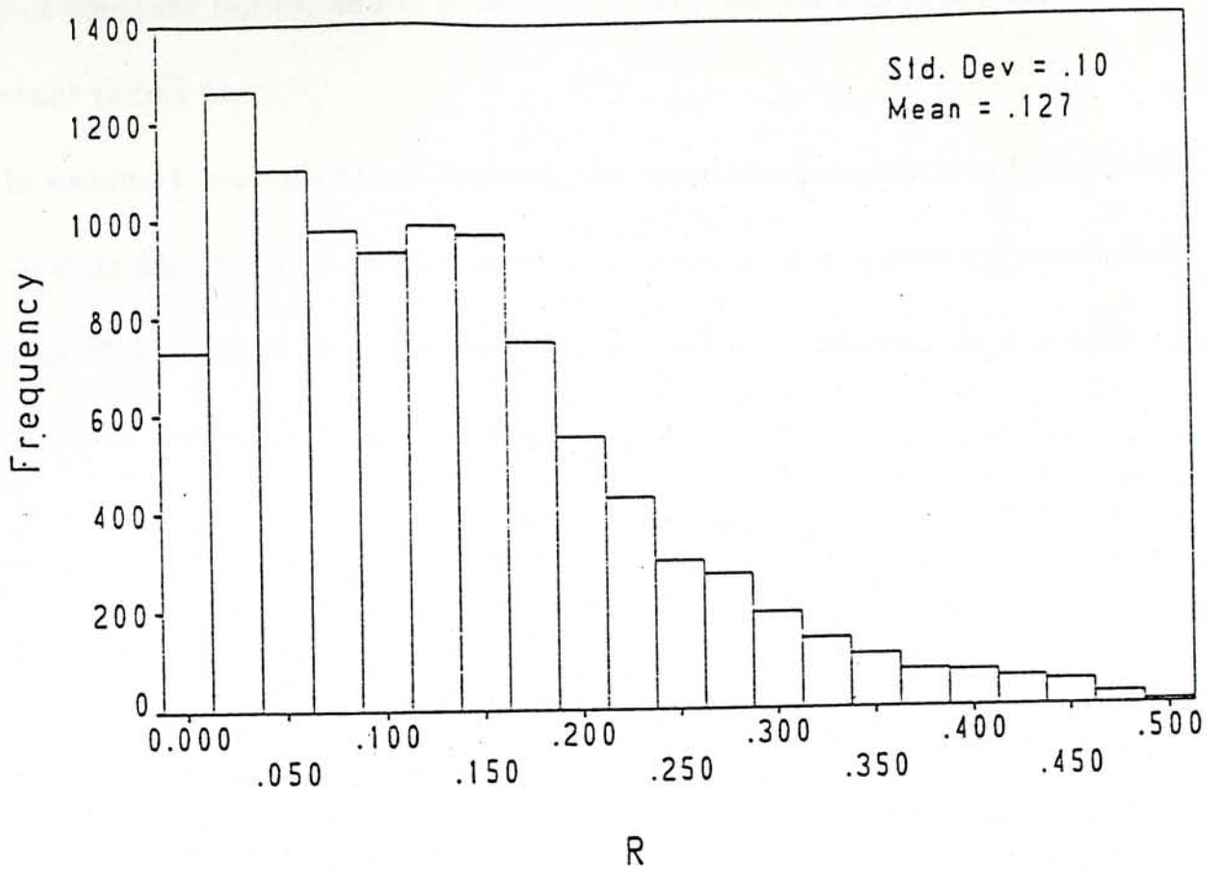

Figure 5.1. The posterior distribution of R with varied radius

In order to estimate the true mean of R, i.e. R_t, we use the mean of the above approximated posterior distribution of R, i.e. $\hat{R}_{bay}$. It is equal to 0.12743 with square error (SE_2), 1.08784E-04. For the independent case, the estimator $\hat{R}_{MME}$, is equal to 0.14898 with square error (SE_1), 1.02243E-03. We find that $\hat{R}_{bay}$ is better than $\hat{R}_{MME}$ in terms of square error with the ratio $(SE_1/SE_2) = 9.3986$.

From the above artifical example, we find that we can extent to the case of varied radius. Besides, it seems to be possible for us to have some further extensions, such as varied radius with N is unknown and the boundary problem. No matter what extensions are, it is more likely to base on the result from the

case of constant radius, and it is the reason why we want to study the effect of constant radius first.

In summary, we notice that imposing the dependence assumption to the traditional stereology problem can improve our estimate of the interesting parameters, such as R or mean of R in the varied radius case. Therefore, it is worth to use the Bayesian estimator instead of the traditional one.

Bibliography

Baddeley, A. (1991) , Stereology, in *Spatial statistics and digital image analysis*, (National Research Council), **181-216**, National Academy Press, Washington D.C.

Casella, G. and George, E.I. (1992). Explaining the Gibbs Sampler, *The American Statistician*, **46**, 3, 167-174.

Diggle, P.J. (1983). *Statistical analysis of spatial point pattern*, Academic Press, New York.

IMSL (1991) STAT/LIBRARY, Version 2.0.

Jensen, E.B. (1987). Design- and model-based stereological analysis of arbitrarily shaped particles, *Scand. J. Statist.*, **14**, 161-180.

Keiding, N. and Andersen, L. (1992). Estimation of the size distribution of fibrillar centres in nucleoli- an example of the "Swiss Cheese" problem in stereology, *Biometrics*, **48**, 449-458.

Keiding, N. and Jensen, S.T. (1972). Maximum likelihood estimation of the size distribution of liver cell nuclei from the observed distribution in a plane

section, *Biometrics*, **28**, 813-929.

Mathematics Handbook (1979). Advanced Education Publishing Co. (In Chinese).

Ripley, B.D. (1979). Algorithm AS 137 : Simulating Spatial Patterns : Dependent samples from a multivariate density, *Applied Statistics*, **28**, 109-112.

Underwood, E.E. (1970). *Quantitative Stereology*, Addison - Wesley Pub. Co., Reading, Mass.

Weibel, E.R. (1979). *Stereology Methods, Vol 1: Practical Method for Biological Morphometry*, Academic Press, London.

Weibel, E.R. (1979). *Stereology Methods, Vol 2: Theoretical Foundations*, Academic Press, London.

Wicksell, S.D. (1925). The corpuscle problem: A mathematical study of a biometrics problem, *Biometrika*, **17**, 84-89.